

NEYMAN'S $C(\alpha)$ TEST FOR UNOBSERVED HETEROGENEITY

JIAYING GU
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

ABSTRACT. A unified framework is proposed for tests of unobserved heterogeneity in parametric statistical models based on Neyman's $C(\alpha)$ approach. Such tests are irregular in the sense that the first order derivative of the log likelihood with respect to the heterogeneity parameter is identically zero, and consequently the conventional Fisher information about the parameter is zero. Nevertheless, local asymptotic optimality of the $C(\alpha)$ tests can be established via LeCam's differentiability in quadratic mean approach. This leads to local alternatives of order $n^{-1/4}$. Many such tests are already familiar from existing literature, but the new framework reveals that certain regularity conditions commonly employed in earlier developments are unnecessary.

1. INTRODUCTION

Neyman's (1959) $C(\alpha)$ test can be viewed as a generalization of Rao's (1948) score test in the presence of nuisance parameters and thus provides a unified framework for parametric statistical inference. We will see that many of the existing tests for neglected parameter heterogeneity can also be formulated as $C(\alpha)$ tests and share common features. However, for these tests the usual score function is identically zero under the null hypothesis, and conventional Fisher information is thus zero. Fortunately, in these cases the second derivative of the log likelihood is non-degenerate and approximations based on it can be used to form a modified version of LeCam's differentiability in quadratic mean (DQM) condition. Local asymptotic normality (LAN) theory, then leads to local asymptotic optimality results for the $C(\alpha)$ test in such settings under local alternatives of order $n^{-1/4}$.

We focus initially on the case of a scalar heterogeneity parameter; extensions to multivariate settings are briefly described at the end of Section 2. In Section 3 we consider three different examples and show that the $C(\alpha)$ test leads to familiar test statistics proposed in the econometric literature for different parametric models. The $C(\alpha)$ tests for parameter heterogeneity in Poisson regression model under two slightly different alternative specifications lead to those in Lee (1986). Kiefer (1984) and Lancaster (1985) derives test for parametric heterogeneity in Cox proportional hazard model which can both be formulated

Date: February 4, 2013.

Department of Economics, University of Illinois at Urbana-Champaign, 214 David Kinley Hall, 1407 W. Gregory Dr., Urbana, Illinois 61801, MC-707, USA. Tel: +1-267-994-1519. Fax: +1-217-244-6571. Email Address: gu17@illinois.edu. I would like to thank Roger Koenker for his continued support and encouragement. I would also like to thank Andreas Hagemann and Stanislav Volgushev for comments and useful discussion. I gratefully acknowledge financial support from NSF grant SES-11-53548 and the Paul Boltz summer Fellowship. All errors are my own.

as $C(\alpha)$ tests. We also construct a $C(\alpha)$ joint test in Gaussian panel data model for heterogeneous location and scale parameter.

The $C(\alpha)$ test for heterogeneity formulated in this paper is very similar to the setup used in some previous development. In a seminal paper, Chesher (1984) points out the score test for unobserved parametric heterogeneity is identical to White's (1982) Information Matrix (IM) test. Cox (1983) obtains similar results using a more general mixture model. These papers can be viewed as important further development to a somewhat neglected example on testing for parameter heterogeneity in a Poisson model in Neyman and Scott (1966). Moran (1973) investigates the asymptotic behavior of these score tests. However, as we will show in Section 4, the parameterization adopted in Moran (1973) and also Chesher (1984) requires unnecessary additional assumptions, even though it delivers the same test statistics as the $C(\alpha)$ test constructed here. We conclude in Section 4 that the $C(\alpha)$ test for unobserved heterogeneity is not always identical to the IM test, and illustrate some conditions for equivalence to hold.

2. THE $C(\alpha)$ TEST FOR UNOBSERVED PARAMETER HETEROGENEITY

Neyman (1959) introduces the $C(\alpha)$ test with the consideration that hypotheses testing problems in applied research often involve several nuisance parameters. In these composite testing problems, most powerful tests do not exist, motivating search for an optimal test procedure that yields the highest power among the class of tests obtaining the same size. The locally asymptotically optimal $C(\alpha)$ test employs regularity conditions inherited from the conditions used by Cramér (1946) for showing consistency of MLE and some further restrictions on the testing function to allow for replacing the unknown nuisance parameters by its \sqrt{n} -consistent estimators. It is the confluence of these Cramér conditions and the maintained significance level α that gives the name to the $C(\alpha)$ test.

2.1. $C(\alpha)$ test in regular cases. In regular cases, where all the score functions with respect to parameters in the model are non-degenerate and the Fisher information matrix is non-singular, the $C(\alpha)$ test is constructed as follows. Suppose we have X_1, \dots, X_n as i.i.d. random variables with density $p(x; \xi, \theta)$ where θ are nuisance parameters belonging to $\Theta \subset \mathbb{R}^p$ and ξ are parameters under test that belong to $\Xi \subset \mathbb{R}^q$. For densities satisfying the regularity conditions (Neyman (1959, Definition 3)), we consider testing the hypothesis $H_0 : \xi = \xi_0$ against $H_a : \xi \in \Xi \setminus \{\xi_0\}$ while nuisance parameters $\theta \in \Theta$ are left unspecified. We define the conventional score functions as

$$C_{\xi, n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\xi} \log p(X_i; \xi, \theta)|_{\xi=\xi_0}$$

$$C_{\theta, n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla_{\theta} \log p(X_i; \xi, \theta)|_{\xi=\xi_0}$$

and define the corresponding matrix of second-order derivatives,

$$I = \begin{pmatrix} I_{\xi\xi} & I_{\xi\theta} \\ I_{\theta\xi} & I_{\theta\theta} \end{pmatrix},$$

as its Fisher information covariance matrix.

Since nuisance parameters θ are left unspecified by H_0 , Neyman (1959) shows that for the test statistic to have the same asymptotic behavior when we replace the nuisance parameters θ by any \sqrt{n} -consistent estimator $\hat{\theta}_n$, it is necessary and sufficient for the test statistics to be orthogonal to $C_{\theta,n}$. For example, the "residual" score, which constitutes the vector of projecting $C_{\xi,n}$ onto the space spanned by the score vector $C_{\theta,n}$, denoted by

$$g_n(\theta) = C_{\xi,n} - I_{\xi\theta} I_{\theta\theta}^{-1} C_{\theta,n},$$

provides such a test function with variance $I_{\xi,\theta} \equiv I_{\xi\xi} - I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi}$. Given a \sqrt{n} -consistent estimator $\hat{\theta}_n$ for θ , the $C(\alpha)$ test

$$T_n(\hat{\theta}_n) = g_n(\hat{\theta}_n)^\top I_{\xi,\theta}^{-1} g_n(\hat{\theta}_n)$$

is then asymptotically χ_q^2 under H_0 and is optimal for local alternatives of the form $\xi_n = \xi_0 + \delta/\sqrt{n}$. When $\hat{\theta}_n$ is the restricted maximum likelihood estimator of θ , $C_{\theta,n}$ is zero and the $C(\alpha)$ test reduces to Rao's score test. The component $I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi}$ subtracted from the information $I_{\xi\xi}$ for ξ measures the amount of information lost due to not knowing the nuisance parameters (see e.g. Bickel, Klaassen, Ritov, and Wellner (1993), section 2.4).

2.2. Testing for unobserved parameter heterogeneity. The $C(\alpha)$ test for unobserved heterogeneity is usually formulated under a random parameter model. Following Neyman and Scott (1966) we will focus initially on testing homogeneity of a scalar parameter against the alternative that the parameter is random. Consider having i.i.d. random variables X_1, \dots, X_n , with each X_i having density function $p(x; \lambda_i)$. Heterogeneity of the model is introduced by regarding the individual specific λ_i as a random parameter of the form,

$$\lambda_i = \lambda_0 + \tau \xi U_i,$$

where the unobserved U_i 's are independent random variables with common distribution function, F , satisfying moment conditions $\mathbb{E}(U) = 0$, $\mathbb{V}(U) = 1$. The parameter τ is a known finite scale parameter. It is not restrictive to assume τ known, as we will see later that τ does not enter the test statistics. The hypothesis we would like to test is $H_0 : \xi = 0$, which implies $\lambda_i = \lambda_0$ for all i 's. The alternative hypothesis is $H_a : \xi \neq 0$.

Under the above setup, the standard $C(\alpha)$ test described in section 2.1 breaks down because the score function for ξ for each individual observation x_i , defined as the first order logarithmic derivative of the density function with respect to ξ , is identically zero under the null, hence the Fisher information is also zero,

$$\frac{\partial}{\partial \xi} \log \int p(x_i; \lambda_0 + \tau \xi u) dF(u) |_{\xi=0} = \tau \int u dF(u) \frac{p'(x_i; \lambda_0)}{p(x_i; \lambda_0)} = 0.$$

However, Neyman (1959, p.224, Corollary 2) was already aware of this possibility and suggested computing the second-order derivative, denoted as $s_i(\lambda_0)$ below,

$$s_i(\lambda_0) := \frac{\partial^2}{\partial \xi^2} \log \int p(x_i; \lambda_0 + \tau \xi u) dF(u) |_{\xi=0} = \tau^2 \int u^2 dF(u) \frac{p''(x_i; \lambda_0)}{p(x_i; \lambda_0)} = \tau^2 \frac{p''(x_i; \lambda_0)}{p(x_i; \lambda_0)}.$$

The normed sum of these independent second-order derivatives, $s(\lambda_0) = \frac{1}{\sqrt{n}} \sum_i s_i(\lambda_0)$, can be shown to be asymptotically normally distributed with mean zero and variance $\mathbb{E}(s_1^2(\lambda_0))$ under H_0 by the central limit theorem. This leads to a close analogy with the regular theorem, in which $s(\lambda_0)$ acts as the score function and the variance $\mathbb{E}(s_1^2(\lambda_0))$ plays the role of the Fisher information in the irregular setting considered here.

In regular cases, score tests exploit the fact that if the null hypothesis is false, the first-order gradient of the log likelihood should not be close to zero. Apparently this fails in the irregular case, because no matter how data is generated, the gradient is always zero. It is natural then to make use of the curvature information (i.e. the second-order condition), provided by the second-order derivative for inference. If the null is false, one expects the second-order derivative to be positive. We will see that this second-order score function plays the essential role of constructing the $C(\alpha)$ test for unobserved heterogeneity. The positivity condition also anticipates the $C(\alpha)$ test to be one-sided. The goal of the remaining part of this section is to show that the optimality of the $C(\alpha)$ test, as in the regular case, is still preserved under this irregularity and its asymptotic theory, although different from the regular cases in certain perspectives, still takes a simple form.

2.3. Asymptotic optimality of the $C(\alpha)$ test for parameter heterogeneity. Under the irregularity discussed above, in order to establish the optimality of the test statistics based on the second-order score function, one could consider modifying the Cramér type regularity conditions in Neyman (1959, Definition 3), requiring the density function to be five times differentiable and impose a Lipschitz condition on the fifth order derivative with respect to the parameter under test. The main motivation is to obtain a quadratic approximation of the log likelihood ratio using the second-order score function through a higher order Taylor expansion. To be more specific, using the example in Section 2.2 as an illustration, for local alternatives $\lambda_i = \lambda_0 + \tau \xi_n \mathbf{U}_i$, with ξ_n be a sequence that converges to zero at certain rate, we have the following Taylor expansion of the log likelihood ratio,

$$\begin{aligned} \Lambda_n = \sum_i \log \frac{p(\mathbf{x}_i; \lambda_i)}{p(\mathbf{x}_i; \lambda_0)} &= \frac{\xi_n^2 \tau^2}{2!} \mathbb{E}(\mathbf{U}^2) \sum_i s_i(\lambda_0) + \frac{\xi_n^3 \tau^3}{3!} \mathbb{E}(\mathbf{U}^3) \sum_i \frac{\nabla_{\lambda}^3 p(\mathbf{x}_i; \lambda_0)}{p(\mathbf{x}_i; \lambda_0)} \\ &+ \frac{\xi_n^4 \tau^4}{4!} \left[\mathbb{E}(\mathbf{U}^4) \sum_i \frac{\nabla_{\lambda}^4 p(\mathbf{x}_i; \lambda_0)}{p(\mathbf{x}_i; \lambda_0)} - 3\mathbb{E}(\mathbf{U}^2)^2 \sum_i s_i^2(\lambda_0) \right] + o_P(1). \end{aligned}$$

Let ξ_n be of order $n^{-1/4}$ and provided the third and fourth moments of \mathbf{U} are finite in addition to the zero mean and unit variance assumption, we obtain a quadratic approximation of the log-likelihood. More details of such regularity conditions can be found in Rotnitzky, Cox, Bottai, and Robins (2000), in which they consider the maximum likelihood estimation of ξ in the irregular cases in a very general context.

An alternative formulation, rooted in LeCam's local asymptotic normality (LAN) theory, can be based on his differentiability in quadratic mean (DQM) condition. The latter condition is less stringent in regular cases: while Cramér conditions assume the density to be three times differentiable and impose a Lipschitz condition on the third order derivative, the DQM condition only requires first order differentiability and the derivative to be square integrable in \mathcal{L}_2 space. Pollard (1997) provides a nice discussion of the DQM condition in

these regular cases. This is the approach we take for analyzing the asymptotic behavior of the $C(\alpha)$ test for heterogeneity. We will show below that by modifying the DQM condition slightly, we can obtain the local asymptotic normality of the log-likelihood ratio and establish the asymptotic optimality of the $C(\alpha)$ test for the irregular cases under assumptions weaker than those suggested by the classical Neyman's approach.

Suppose we have a random sample (X_1, \dots, X_n) with density function $p(x; \xi, \theta)$ with respect to some measure μ . The joint distribution of this i.i.d. random sample will be denoted as $P_{n, \xi, \theta}$, which is the product of n copies of the marginal distribution $P(x; \xi, \theta)$.

Assumption 1. The density function p satisfies the following conditions:

- (1) ξ_0 is an interior point of Ξ
- (2) For all $\theta \in \Theta \subset \mathbb{R}^p$ and $\xi \in \Xi \subset \mathbb{R}$, the density is twice continuously differentiable with respect to ξ and once continuously differentiable with respect to θ for all x .
- (3) Denoting the first two derivatives of the density with respect to ξ evaluated under the null as $\nabla_\xi p(x; \xi_0, \theta)$ and $\nabla_\xi^2 p(x; \xi_0, \theta)$, we have $\mathbb{P}(\nabla_\xi p(x; \xi_0, \theta) = 0) = 1$ and $\mathbb{P}(\nabla_\xi^2 p(x; \xi_0, \theta) \neq 0) > 0$.
- (4) Denoting the derivative of the density with respect to θ evaluated under the null as $\nabla_\theta p(x; \xi_0, \theta)$, for any p -dimensional vector \mathbf{a} , $\mathbb{P}(\nabla_\xi^2 p(x; \xi_0, \theta) \neq \mathbf{a}^\top \nabla_\theta p(x; \xi_0, \theta)) > 0$.

Remark. Here ξ is the parameter under test and θ is the vector of nuisance parameters. The list of regularity conditions in Assumption 1 tailors the standard conditions for a regular $C(\alpha)$ test to the heterogeneity test we consider here. In particular, condition (3) reflects the irregularity of these tests that the first order logarithmic derivative with respect to ξ vanishes but the second-order derivative is non-vanishing. Condition (2) secures existence of the respective derivatives. Condition (4) rules out the case where there is a perfect linear relationship between the second-order score for ξ and the score for θ . It ensures the new Fisher information thus defined to be non-singular and the $C(\alpha)$ test statistics to be non-degenerate.

Under Assumption 1, we can now define the modified DQM condition that is crucial for establishing the local asymptotic normality of the model.

Definition 1. The density $p(x; \xi, \theta)$ satisfies the modified differentiability in quadratic mean condition at (ξ_0, θ) if there exists a vector $\mathbf{v}(x) = (\mathbf{v}_\xi(x), \mathbf{v}_\theta^\top(x))^\top \in \mathcal{L}_2(\mu)$ such that as $(\xi_n, \theta_n) \rightarrow (\xi_0, \theta)$,

$$\int |\sqrt{p(x; \xi_n, \theta_n)} - \sqrt{p(x; \xi_0, \theta)} - \mathbf{h}_n^\top \mathbf{v}(x)|^2 d\mu(x) = o(\|\mathbf{h}_n\|^2)$$

where $\mathbf{h}_n = ((\xi_n - \xi_0)^2, (\theta_n - \theta)^\top)^\top$. Here $\|\cdot\|$ denotes the Euclidean norm and $\mathcal{L}_2(\mu)$ denotes the \mathcal{L}_2 space of square integrable functions with respect to measure μ .

Furthermore, let $\beta(\mathbf{h}_n)$ be the mass of the part of $p(x; \xi_n, \theta_n)$ that is $p(x; \xi_0, \theta)$ -singular,

then as $(\xi_n, \theta_n) \rightarrow (\xi_0, \theta)$,

$$\frac{\beta(\mathbf{h}_n)}{\|\mathbf{h}_n\|^2} \rightarrow 0$$

Usually the vector $\mathbf{v}(\mathbf{x})$ contains derivatives of the square root of density $\sqrt{p(\mathbf{x}; \xi_n, \theta_n)}$ with respect to each parameters evaluated under their null value. Definition 1 modifies the classical DQM condition such that whenever the first order derivative is identically zero for certain parameters, it is differentiated again until it is nonvanishing. The corresponding terms in \mathbf{h}_n also need to be raised to the same power. For the heterogeneity test, the score function with respect to ξ is of second order and its associated term in \mathbf{h}_n is hence quadratic. This further implies that the contiguous alternatives must be $O(n^{-1/4})$. For the following theorems, we will thus focus on the sequence of local models on (X_1, \dots, X_n) with joint distribution P_{n, ξ_n, θ_n} in which $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and $\theta_n = \theta + \delta_2 n^{-1/2}$.

Theorem 1. Suppose (X_1, \dots, X_n) are i.i.d. random variables with joint distribution P_{n, ξ_n, θ_n} and the density satisfies Assumption 1 and the modified DQM condition with

$$\mathbf{v}(\mathbf{x}) = (\mathbf{v}_\xi(\mathbf{x}), \mathbf{v}_\theta^\top(\mathbf{x}))^\top = \left(\frac{1}{4} \frac{\nabla_\xi^2 p(\mathbf{x}; \xi_0, \theta)}{\sqrt{p(\mathbf{x}; \xi_0, \theta)}} \mathbb{I}_{[p(\mathbf{x}; \xi_0, \theta) > 0]}, \frac{1}{2} \frac{\nabla_\theta p(\mathbf{x}; \xi_0, \theta)}{\sqrt{p(\mathbf{x}; \xi_0, \theta)}} \mathbb{I}_{[p(\mathbf{x}; \xi_0, \theta) > 0]} \right)^\top,$$

then for fixed δ_1 and δ_2 , the log-likelihood ratio has the following quadratic approximation under the null:

$$\Lambda_n = \log \frac{dP_{n, \xi_n, \theta_n}}{dP_{n, \xi_0, \theta}} = \mathbf{t}^\top S_n - \frac{1}{2} \mathbf{t}^\top J \mathbf{t} + o_P(1)$$

where $\mathbf{t} = (\delta_1^2, \delta_2^\top)^\top$,

$$S_n = (S_{\xi, n}, S_{\theta, n}^\top)^\top = \left(\frac{2}{\sqrt{n}} \sum_i \frac{\mathbf{v}_\xi(\mathbf{x}_i)}{\sqrt{p(\mathbf{x}_i; \xi_0, \theta)}}, \frac{2}{\sqrt{n}} \sum_i \frac{\mathbf{v}_\theta^\top(\mathbf{x}_i)}{\sqrt{p(\mathbf{x}_i; \xi_0, \theta)}} \right)^\top$$

and

$$J = 4 \int (\mathbf{v} \mathbf{v}^\top) d\mu(\mathbf{x}) = \begin{pmatrix} \mathbb{E}(S_{\xi, n}^2) & \text{Cov}(S_{\xi, n}, S_{\theta, n}^\top) \\ \text{Cov}(S_{\xi, n}, S_{\theta, n}) & \mathbb{E}(S_{\theta, n} S_{\theta, n}^\top) \end{pmatrix} \equiv \begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix}.$$

Corollary 1. With S_n and J defined as in Theorem 1, we have

$$S_n \overset{P_{n, \xi_0, \theta}}{\rightsquigarrow} \mathcal{N}(0, J),$$

and hence the sequence of models P_{n, ξ_n, θ_n} is locally asymptotically normal (LAN) at (ξ_0, θ) with S_n being interpreted as the score vector and J as the associated Fisher information matrix. Furthermore, P_{n, ξ_n, θ_n} is mutually contiguous to $P_{n, \xi_0, \theta}$.

Theorem 1 shows that under Assumption 1, the modified DQM condition is sufficient for obtaining a quadratic approximation of the log-likelihood ratio for the sequence of local models in the $n^{-1/4}$ neighborhood of the null value ξ_0 and the $n^{-1/2}$ neighborhood of the nuisance parameter θ . The joint normality of the vector S_n , as established in Corollary 1, further indicates the LAN property of this sequence of models. It is important to note that the vector S_n , in which the degenerately zero first-order score function for ξ is replaced by the corresponding second-order derivative of the log-likelihood, acts as the score vector in this irregular case. Naturally, J has the interpretation of the Fisher information matrix. Under Assumption 1, since we rule out perfect dependence between $S_{\xi,n}$ and $S_{\theta,n}$ in condition (4), J is non-singular.

Having established the LAN property of this sequence of local models, we can now make use of LeCam's (1972) limit experiment theory to show that the $C(\alpha)$ test is locally asymptotically optimal.

Following the definitions given in LeCam (1972) and van der Vaart (1998), an experiment \mathcal{E} indexed by a parameter set H is a collection of probability measures $\{P_h : h \in H\}$ on the sample space $(\mathcal{X}, \mathcal{A})$. A sequence of experiments $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ is said to converge to a limit experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$ if the likelihood ratio process for \mathcal{E}_n , $\frac{dP_{n,h}}{dP_{n,h_0}}(X_n)$, converges in distribution to the likelihood ratio of the limit experiment, $\frac{dP_h}{dP_{h_0}}(X)$, for h in a finite subset $I \subset H$ and h_0 being the true null value. A common feature is that many sequence of experiments obtain a Gaussian limit experiment. One important example is that for i.i.d. sample from a smooth parametric model with distribution P_ϑ , if the sequence of the local model P_{n,ϑ_n} in which $\vartheta_n = \vartheta_0 + r_n \delta$ with r_n as the appropriate normalizing rate is locally asymptotically normal, then it has a Gaussian shift experiment as its limit.

The advantage of establishing the limit experiment is several fold. First, the limit experiment is often easier to analyze than the original sequence of models. Second, the limit experiment provides a bound for the optimal estimation (in terms of lower bound on the asymptotic variance) or testing procedure (in terms of upper bound on the asymptotic power) one could achieve in the original model. Third, by van der Vaart's (1991)'s asymptotic representation theory, the optimal procedure found for the limit experiment can be matched to a sequence of statistics in the original experiment and they preserve the identical asymptotic behavior. We will show below that the optimal test statistic found in the Gaussian shift limit experiment is matched by the $C(\alpha)$ test for the heterogeneity test problems. We will first focus on the scalar case, leaving the multi-dimensional case to a separate discussion.

Theorem 2. Let \mathcal{E}_n be a sequence of experiments based on i.i.d. random variables (X_1, \dots, X_n) with joint distribution P_{n,ξ_n,θ_n} on the sample space $(\mathcal{X}_n, \mathcal{A}_n)$. We further index the sequence of experiment by $t = (\delta_1^2, \delta_2^2)^\top \in \mathbb{R}_+ \times \mathbb{R}^p$. The log-likelihood ratio of the sequence of models satisfies,

$$\log \left(\frac{dP_{n,\xi_n,\theta_n}}{dP_{n,\xi_0,\theta}} \right) = t^\top S_n - \frac{1}{2} t^\top J t + o_P(1),$$

with the score vector S_n defined as in Theorem 1 converging in distribution under the null to $N(0, J)$. Then the sequence of experiments \mathcal{E}_n converges to the limit experiment based on observing one sample from $Y = t + v$, where $v \sim N(0, J^{-1})$. The locally asymptotically optimal statistic for testing $H_0 : \delta_1 = 0$ vs. $H_a : \delta_1 \neq 0$ is

$$Z_n = (J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{-1/2}(S_{\xi,n} - J_{\xi\theta}J_{\theta\theta}^{-1}S_{\theta,n}).$$

Corollary 2. Under H_0 , Z_n has distribution $N(0, 1)$. Under H_a , by applying LeCam's third lemma (see e.g. van der Vaart (1998, Example 6.7)), it follows a shifted normal distribution $N(\delta_1^2(J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{1/2}, 1)$.

The optimal test statistic Z_n takes the form of a $C(\alpha)$ test. It projects the second-order score $S_{\xi,n}$ for ξ onto the space spanned by the first-order score vector $S_{\theta,n}$ for θ . It is the sequence of statistics from the original experiment matched to the optimal test statistic in the limit Gaussian experiment for inference on δ_1 .

One common feature of $C(\alpha)$ heterogeneity tests is that the limit distribution under local alternative is always a right-shifted normal distribution even if we have a two-sided alternative hypothesis for δ_1 . This is not surprising given that the shift parameter corresponding to ξ in the Gaussian limit experiment is a quadratic term $\delta_1^2 \in \mathbb{R}_+$. In other words, the best inference procedure one could possibly achieve in the limit experiment is for δ_1^2 . We lose the sign information on δ_1 , and the asymptotically optimal test, if rejects the null, fails to distinguish whether the deviation is from the left or from the right (this phenomenon is also accentuated in Rotnitzky, Cox, Bottai, and Robins (2000)). The one-sidedness of the test implies that we reject H_0 if $(0 \vee Z_n)^2 > c$, where c is the $(1 - \alpha)$ -quantile of $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ and χ_0^2 is a degenerate distribution with mass 1 at 0. The weight 1/2 associated with χ_0^2 is due to the fact that Z_n takes negative value with probability 1/2 under H_0 .

There is another intuitive interpretation of the one-sidedness of the test, as we have already anticipated in Section 2.2. The $C(\alpha)$ test Z_n , constructed from the second-order score for ξ , exploits information of the curvature of the log-likelihood function. Since at $\xi = \xi_0$, the gradient of the log-likelihood function with respect to ξ is always zero, it depends on the sign of the second-order derivative to determine whether the null point is a local maxima or a local minima. Only positive value of Z_n indicates the null point as a local minima of the log-likelihood function, leading to a rejection of the null hypothesis. As $n \rightarrow \infty$, due to normality of Z_n , only half the time we get the "correct" curvature allowing us to reject the null.

In our random parameter model, one could of course also consider a likelihood ratio test as an alternative testing strategy for heterogeneity. Among many others, Chen, Chen, and Kalbfleisch (2001) considers a modified likelihood ratio test for homogeneity in finite mixture models, which is very close to the setup we consider in this paper. They also obtain a mixture of χ^2 asymptotics for their likelihood ratio test statistics. The modified LRT can be viewed as an asymptotically equivalent testing procedure in finite mixture models to the $C(\alpha)$ test considered here. The latter, however, inheriting the nice feature of the score test, is much easier to compute. Furthermore, the $C(\alpha)$ test statistics does not depend on the

specification of F as long as the moment conditions are satisfied. This can be viewed as a merit of the test because it has power for a large class of alternative models. On the other hand, it can also be viewed as its disadvantage because rejecting the hypothesis does not provide information on what plausible alternative might be.

The result established thus far is not specialized to the heterogeneity test problem. It is applicable whenever the first-order score for the parameter under test vanishes but the second-order score is non-degenerate. There is another possible scenario for the score test to break down, in which none of the first-order score function is vanishing, but there is linear dependence among them, and thus the Fisher information matrix becomes singular. This is the case discussed in much details in Lee and Chesher (1986). Models with selection bias and the stochastic production frontier models fall into this class. They propose an extreme test which is based on the determinant of the matrix of the second-order derivatives of the log likelihood function and show the asymptotic optimality of the test. The extremum test can essentially be reformulated, using a reparameterization slightly different from what the authors suggested in the paper (i.e. choose k to be 1 in Lee and Chesher (1986, p. 132)), to fit into the conditions described in Assumption 1. The similar irregularity also arises in test for symmetry in normal-skew distribution and is investigated in Hallen and Ley (2012). The reparameterization is a Gram-Schmidt orthogonalization in the same spirit of Rotnitzky, Cox, Bottai, and Robins (2000, Section 4.4)). The $C(\alpha)$ test can then be constructed and asymptotic optimality of the test follows.

2.4. Replacing the nuisance parameter by a \sqrt{n} -consistent estimator. Notice that the optimal test statistic Z_n we obtained in Theorem 2 is a function of θ , to make the test statistic feasible under unknown nuisance parameters, we need to replace θ by some estimator $\hat{\theta}$. In order to ensure that the asymptotics for the test statistic Z_n in Corollary 2 is still valid, it suffices to show that $Z_n(\hat{\theta}) - Z_n(\theta) = o_p(1)$ both under the null and local alternatives. There are various ways to obtain this result. The classical approach taken in Neyman (1959) was to make additional differentiability and bound conditions on the test function $g(x_i, \theta)$, which is defined as

$$g(x_i; \theta) = (J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{-1/2} \left(\frac{2v_{\xi}(x_i)}{\sqrt{p(x_i; \xi_0, \theta)}} - J_{\xi\theta}J_{\theta\theta}^{-1} \frac{2v_{\theta}(x_i)}{\sqrt{p(x_i; \xi_0, \theta)}} \right),$$

such that $Z_n(\theta) = \frac{1}{\sqrt{n}} \sum_i g(x_i, \theta)$. Details of these assumptions can be found in Neyman (1959, Definition 3 (ii) (iii)) and we will not replicate them here. When the conditions are satisfied, Taylor expansion of $Z_n(\hat{\theta})$ around $Z_n(\theta)$ yields the desired results for $\hat{\theta}$ being any \sqrt{n} -consistent estimator for θ . Neyman's assumptions are rather strong, for example, he requires the density to be three times differentiable with respect to θ and also moments of the gradient of g with respect to θ to be continuous. Another approach, using more modern probability theory, is to view the difference $Z_n(\hat{\theta}) - Z_n(\theta)$ as an empirical process. More precisely, we make the following assumption on the test function $g(x, \theta)$.

Assumption 2. There exists some $\delta > 0$ such that for any $\eta, \eta' \in \mathcal{U}_{\delta}(\theta)$ we have for some $\gamma > 0$

$$|g(x, \eta) - g(x, \eta')| \leq \|\eta - \eta'\|^{\gamma} H(x)$$

for $P_{n,\xi_n,\theta}$ -almost all x (for every $n \in \mathbb{N}$) where H is square integrable with respect to $P_{n,\xi_n,\theta}$ for all $n \in \mathbb{N}$, $\sup_n \mathbb{E}_{P_{n,\xi_n,\theta}} H^2(X) < \infty$ and additionally for some $c_n = o(1)$, $n^{1/2} \mathbb{E}_{P_{n,\xi_n,\theta}} [H(X) \mathbb{I}_{\{H(X) > n^{1/2} c_n\}}] = o(1)$.

Theorem 3. Under Assumption 2, if $\hat{\theta}$ is a \sqrt{n} -consistent estimator for θ , then

$$|Z_n(\hat{\theta}) - Z_n(\theta)| = o_P(1)$$

2.5. $C(\alpha)$ test for parameter heterogeneity in higher dimensions. It is of interest to generalize the $C(\alpha)$ tests of unobserved parameter heterogeneity to higher dimensions in the irregularity case. For example, under a Gaussian model, we may want to jointly test for heterogeneity in both location and scale parameters. The main challenge comes from the one-sidedness of the test. In higher dimensions, it is natural to look for analogues of a one-sided test. The limit experiment turns out to be multivariate Gaussian with location shifts in each coordinates towards the right tail. This requires us to look for optimal tests for deviations of the location parameters from zero restrictions to the positive orthant.

To be more specific, suppose the limit multivariate Gaussian experiment has mean vector (μ_1, \dots, μ_q) , we would like to test $H_0 : \mu_i = 0$ for $i = 1, \dots, q$ against the alternative $H_a : \mu_i \geq 0$ for $i = 1, \dots, q$ with at least one inequality holds strictly. This testing problem has been studied by several authors, particularly for the likelihood ratio test. Chernoff (1954) extends the classical Wilks's result on likelihood ratio test (LRT) to cases in which the null value of the parameters under test lie on the boundary of the parameter space. Hillier (1986) provides details for the LRT with dimension equal to three. Self and Liang (1987) give some further examples for LRT with nuisance parameters. Gouriéroux, Holly, and Monfort (1982) considers testing problems in linear regression models with non-negative constraints on the regression coefficients. Test statistics for these one-sided test problems in multi-dimensions all obtain a mixture of χ^2 with different degrees of freedom as their asymptotic distribution. The weights of these χ^2 's get complicated very quickly as dimension increases. We will present in detail the joint test for heterogeneity in two parameters as an illustration and comment on the more general case. Bühlér and Puri (1966) extends the regular $C(\alpha)$ case to higher dimensions. The asymptotic distribution for multidimensional $C(\alpha)$ test for heterogeneity is different from that in Bühlér and Puri (1966) due to the positivity constraints.

Suppose again we have i.i.d. random sample (X_1, \dots, X_n) with density $p(x; \xi, \theta)$. The parameters under test are now $\xi = (\xi_1, \xi_2) \in \Xi \subset \mathbb{R}^2$. They take null value $\xi_0 = (\xi_{10}, \xi_{20})$ and $\theta \in \Theta \subset \mathbb{R}^p$ are the nuisance parameters. For heterogeneity tests in particular, we consider testing for heterogeneity of a vector of parameters, λ_i , of the model. Under the alternative, they take the form, $\lambda_{ik} = \theta_k + \tau \xi_k U_{ik}$, for $k = 1, 2$ and U_{ik} are independent random variables. Under $H_0 : \xi_k = 0$, so that λ_k 's are homogenous across individuals taking value θ_k .

The density function satisfies Assumption 1 such that the first-order score for ξ_1 and ξ_2 are vanishing but the second-order score is non-vanishing. It also satisfies the modified DQM

condition so that the model is locally asymptotically normal. We denote the second-order score for (ξ_1, ξ_2) as $(S_{\xi_1, n}, S_{\xi_2, n})$ and the first-order score for θ as $S_{\theta, n}$. More specifically, under regularity conditions, they are $S_{\xi_k, n} = \frac{1}{2\sqrt{n}} \sum_i \frac{\nabla_{\xi_k}^2 p(x_i; \xi_0, \theta)}{p(x_i; \xi_0, \theta)} \mathbb{I}_{[p(x_i; \xi_0, \theta) > 0]}$ and $S_{\theta, n} = \frac{1}{\sqrt{n}} \sum_i \frac{\nabla_{\theta} p(x_i; \xi_0, \theta)}{p(x_i; \xi_0, \theta)} \mathbb{I}_{[p(x_i; \xi_0, \theta) > 0]}$. Let the associated information matrix be denoted as,

$$J = \begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix},$$

with $J_{\xi\xi}$ being a 2×2 block matrix. The residual score for ξ , similar to the scalar case, is found to be

$$\tilde{S}_{\xi, n} = \begin{pmatrix} \tilde{S}_{\xi_1, n} \\ \tilde{S}_{\xi_2, n} \end{pmatrix} := \begin{pmatrix} S_{\xi_1, n} - J_{\xi\theta} J_{\theta\theta}^{-1} S_{\theta, n} \\ S_{\xi_2, n} - J_{\xi\theta} J_{\theta\theta}^{-1} S_{\theta, n} \end{pmatrix}$$

and the covariance matrix for $\tilde{S}_{\xi, n}$ is $\Sigma = J_{\xi\xi} - J_{\xi\theta} J_{\theta\theta}^{-1} J_{\theta\xi} := \begin{pmatrix} v_1 & \sigma_{12} \\ \sigma_{12} & v_2 \end{pmatrix}$. Let Λ be the Cholesky decomposition of matrix Σ , that

$$\Lambda = \begin{pmatrix} \sqrt{v_1} & 0 \\ \rho\sqrt{v_2} & \sqrt{v_2}\sqrt{1-\rho^2} \end{pmatrix}$$

where $\rho = \sigma_{12}/(\sqrt{v_1}\sqrt{v_2})$ is the correlation coefficient between $\tilde{S}_{\xi_1, n}$ and $\tilde{S}_{\xi_2, n}$.

Theorem 4. Let \mathbf{v}_n be the sequence of experiments based on i.i.d. random variable (X_1, \dots, X_n) with joint distribution P_{n, ξ_n, θ_n} with $\xi_n = (\xi_{10}, \xi_{20}) + (\delta_1, \delta_2)n^{-1/4}$ and $\theta_n = \theta + \delta_3 n^{-1/2}$ on the sample space $(\mathcal{X}_n, \mathcal{A}_n)$. The log-likelihood ratio of the sequence of experiment satisfies,

$$\log \left(\frac{dP_{n, \xi_n, \theta_n}}{dP_{n, \xi_0, \theta}} \right) = \mathbf{t}^\top S_n - \frac{1}{2} \mathbf{t}^\top J \mathbf{t} + o_p(1),$$

with $S_n = (S_{\xi_1, n}, S_{\xi_2, n}, S_{\theta, n})^\top \sim \mathcal{N}(0, J)$. Then the limit experiment of \mathbf{v}_n is based on observing one sample from $Y = \mathbf{t} + \mathbf{v}$ with $\mathbf{t} = (\delta_1^2, \delta_2^2, \delta_3^2)^\top \in \mathbb{R}_+^2 \times \mathbb{R}$ and $\mathbf{v} \sim \mathcal{N}(0, J^{-1})$. We would like to jointly test $H_0 : \delta_1 = \delta_2 = 0$ against the alternative $H_a : \delta_1 \neq 0$ or $\delta_2 \neq 0$. Define $w_n = (w_{1n}, w_{2n})$ as

$$w_n \equiv \Lambda^{-1} \tilde{S}_{\xi, n} = \begin{pmatrix} \tilde{S}_{\xi_1, n} / \sqrt{v_1} \\ (1 - \rho^2)^{-1/2} (\tilde{S}_{\xi_2, n} / \sqrt{v_2} - \rho \tilde{S}_{\xi_1, n} / \sqrt{v_1}) \end{pmatrix}$$

The optimal $C(\alpha)$ test statistic is one of the following four cases:

$$T_n = \begin{cases} w_{1n}^2 + w_{2n}^2 & \text{if } w_{1n} \geq \frac{\rho}{\sqrt{1-\rho^2}} w_{2n}, w_{2n} \geq 0 \\ w_{1n}^2 & \text{if } w_{2n} \leq 0, w_{1n} \geq 0 \\ (\rho w_{1n} + \sqrt{1-\rho^2} w_{2n})^2 & \text{if } -\frac{\sqrt{1-\rho^2}}{\rho} w_{2n} \leq w_{1n} \leq \frac{\rho}{\sqrt{1-\rho^2}} w_{2n} \\ 0 & \text{if } w_{1n} \leq 0, w_{2n} \leq -\frac{\rho}{\sqrt{1-\rho^2}} w_{1n} \end{cases}$$

Under H_0 , the asymptotic distribution of T_n follows $(\frac{1}{2} - \frac{\beta}{2\pi})\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{\beta}{2\pi}\chi_2^2$ with $\beta = \cos^{-1}(\rho)$.

Remark. The optimal test statistic T_n is constructed by sequential conditioning. We first find the residual score $\tilde{S}_{\xi,n}$ for (ξ_1, ξ_2) by conditioning on score for θ . LeCam's third lemma implies that asymptotically $\tilde{S}_{\xi,n}$ follows $\mathcal{N}(0, \Sigma)$ under H_0 and $\mathcal{N}((\delta_1^2, \delta_2^2)\Sigma, \Sigma)$ under local alternatives. A further conditioning (See Rao (1973, p.523)) breaks the multivariate normal for $\tilde{S}_{\xi,n}$ into two orthogonal marginals. In particular, w_{2n} is the conditional of $\tilde{S}_{\xi_2,n}$ on $\tilde{S}_{\xi_1,n}$.

As the dimension of parameter under test grows, the sequential conditioning argument described above implies that the optimal test statistics will still follow a mixture of χ^2 distribution asymptotically under the null, albeit with more complex weights. If J happens to be a diagonal matrix, the weights take a very simple form. For $\xi \in \Xi \subset \mathbb{R}^q$ and let the residual score for ξ be $\tilde{S}_{\xi,n}$ with its covariance matrix as Σ_q . The diagonality of J implies diagonality of Σ_q . The optimal test statistic for $H_0 : \xi_1 = \dots = \xi_q = 0$ against $H_a : \xi_i \neq 0$ for at least one i is

$$T_n = (0 \vee \tilde{S}_{\xi,n})^\top \Sigma_q^{-1} (0 \vee \tilde{S}_{\xi,n})$$

Under H_0 , $T_n \sim \sum_{i=0}^q \binom{q}{i} 2^{-q} \chi_i^2$.

3. EXAMPLES

In this section, we describe three examples of using the $C(\alpha)$ test for unobserved parameter heterogeneity in various models. The first two examples lead to similar test statistics already familiar in the literature. The last example on jointly testing for location and scale parameter heterogeneity in Gaussian panel data model is new. Special cases of this example lead to test statistics similar to the test for individual effects in Breusch and Pagan (1980).

3.1. Tests for overdispersion in Poisson Regression. Overdispersion tests for Poisson models constitute the most common example on test of parameter heterogeneity. Such a test was proposed in Fisher (1950) and also serves as the motivating example in Neyman and Scott (1966). We will consider two distinct versions of the test for heterogeneity of the intercept in the Poisson regression model.

3.1.1. Second Moment Test. Suppose we have (Y_1, \dots, Y_n) as i.i.d. random variables follow Poisson distribution with mean parameter λ_i . We further assume that

$$\lambda_i = \lambda_{0i} e^{\xi U_i} = \exp(x_i' \beta + \xi U_i)$$

where U_i are i.i.d. with distribution F , zero mean and unit variance. We have set τ to be 1 without loss of generality. The x_i 's are covariates of the Poisson regression model including an intercept term. These covariates could be viewed as observed heterogeneity of the mean parameter, while U_i , since it is not explained by the covariates, is unobserved heterogeneity. Thus, the intercept coefficient, β_0 , given the assumed form for λ_i , can be regarded as a random coefficient. We would like to test $H_0 : \xi = 0$ against $H_a : \xi \neq 0$ with β as the

unspecified nuisance parameters. Since the first-order score with respect to ξ vanishes, this problem falls into the framework we considered in Section 2.

Let $f(y_i; \xi, \beta)$ be the Poisson density function, with the respective second-order score for ξ and the first-order score for β , the residual score is defined as:

$$g(y_i, \beta) = \nabla_\xi^2 \log f(y_i; 0, \beta) - \mathbf{a}^\top \nabla_\beta \log f(y_i; 0, \beta)$$

where \mathbf{a} is the regression coefficients of projecting the score of ξ on the space spanned by the score vector of β . Under H_0 , it is easy to see that $\mathbf{a}^\top = [1, 0, \dots, 0]$ and hence

$$g(y_i, \beta) = (y_i - \exp(\mathbf{x}'_i \beta))^2 - \exp(\mathbf{x}'_i \beta) - (y_i - \exp(\mathbf{x}'_i \beta))$$

and further, $\mathbb{V}(g(Y_i, \beta)) = 2 \exp(2\mathbf{x}'_i \beta)$. If the MLE $\hat{\beta}$, found by solving the normal equations, $\sum_i (y_i - \exp(\mathbf{x}'_i \hat{\beta})) \mathbf{x}_i = 0$, is used as the \sqrt{n} -consistent estimator to replace β , we have the locally optimal $C(\alpha)$ test statistic as

$$Z_n = \frac{\sum_i g(y_i, \hat{\beta})}{\sqrt{\sum_i \mathbb{V}(g(Y_i, \hat{\beta}))}} = \frac{\sum_i [(y_i - \exp(\mathbf{x}'_i \hat{\beta}))^2 - \exp(\mathbf{x}'_i \hat{\beta})]}{\sqrt{2 \sum_i \exp(2\mathbf{x}'_i \hat{\beta})}}$$

We call this the second moment test because Z_n is essentially comparing the sample second moment with the second moment for the Poisson model under H_0 .

Remark. The $C(\alpha)$ test constructed above is identical to the first test statistic proposed in Lee (1986) for over dispersion in Poisson regression models. It has also been discussed in Collings and Margolin (1985) and Cameron and Trivedi (1986), among many others, and can be viewed as an extension to Fisher's (1950) dispersion test for univariate Poisson models. In his derivation, Lee assumed that the Poisson mean parameter, λ_i , follows a Gamma distribution with certain mean-variance ratio. The Poisson-Gamma compound distribution then leads to a negative binomial model. As Lee noted (p.700), the same test statistic can also be derived under some other distribution in addition to the Gamma distribution (See also Dean and Lawless (1989)). From the $C(\alpha)$ perspective, the test statistic does not depend on the distribution of \mathbf{U} , as long as the moment conditions are satisfied. However, it does depend on the particular specification on λ_i as a function of the observed covariates and the unobservable \mathbf{U}_i . This corresponds to the remark in Lee (1986) that if the mean-variance ratio for the Gamma distribution imposed on λ_i is modified, one arrives at a different test statistic. He denotes it as the second factorial moment test. In the following example, we give the corresponding $C(\alpha)$ formulation.

3.1.2. Second Factorial Moment Test. If instead, under the same setup as we have in 3.1.1, we assume,

$$\lambda_i = \lambda_{0i} \left(1 + \xi \mathbf{U}_i / \sqrt{\lambda_{0i}} \right)$$

The residual score for ξ is now found to be, with $\lambda_{0i} = \exp(\mathbf{x}'_i \beta)$,

$$g(y_i, \beta) = [y_i(y_i - 1) - 2\lambda_{0i}(y_i - \lambda_{0i}) - \lambda_{0i}^2] / \lambda_{0i}$$

and $\mathbb{V}(g(Y_i, \beta)) = 2$. Replacing β by its restricted MLE $\hat{\beta}$, the locally optimal $C(\alpha)$ test is

$$Z_n = \frac{1}{\sqrt{2n}} \sum_i \left[y_i(y_i - 1) - \hat{\lambda}_{0i}^2 \right] / \hat{\lambda}_{0i}$$

This is called the second factorial moment test because Z_n is comparing the second sample factorial moment with that induced by a Poisson model. Note that this test reduces to the second moment test if there are no covariates.

Remark. Cox (1983) and Chesher (1984) provides a general framework of deriving local score test that has power against a general mixed Poisson alternative. Both approaches can be viewed as a $C(\alpha)$ test for a particular form of the random-parameter heterogeneity. Chesher (1984) also discusses an important link between the score test for heterogeneity and the Information Matrix test introduced by White (1982). For the two examples in the Poisson regression model, the Information Matrix test with respect to the intercept term is identical to the second moment test, but not to the second factorial moment test. This leads us to conclude that the $C(\alpha)$ test for heterogeneity is not in general identical to the Information Matrix test if we allow for covariates in the model. We will give a more general discussion on conditions for equivalence to hold between the two in Section 4.

3.2. The Cox Proportional Hazard Model with Frailty. Introducing random effects into survival models is attractive because it is often implausible to make the assumption that individuals are homogenous even when the model includes covariates to control for the observed heterogeneity. Unobserved heterogeneity is often called frailty in survival models, as first proposed by Vaupel, Manton, and Stallard (1979). Survival models play an important role in the economics literature, especially on research for unemployment duration. Lancaster and Nickell (1980) and Heckman and Singer (1982) demonstrate that neglecting time invariant unobserved individual heterogeneity in survival models, i.e. the Cox's proportional hazard model, can lead to wrong inferences on duration dependence. It is therefore of interest to develop a test for frailty in these models.

In the proportional hazard model, without further assumptions on the cumulative hazard function or the distribution of the frailty term, it is impossible to distinguish the effects of duration dependence and unobserved individual heterogeneity. For this reason, researchers often impose rather strong parametric assumptions. For example, Lancaster and Nickell (1980) assume a Weibull distribution for the baseline hazard and introduce a frailty term following Gamma distribution. However, estimation results are usually sensitive to these arbitrary assumptions. Fortunately, Elbers and Ridder (1982) shows that if the unobserved frailty term is multiplicative on the hazard and it has finite mean, with sufficient variation in covariates x , it is possible to nonparametrically identify the model. Heckman and Singer (1984) replace the finite mean assumption by a tail restriction on the frailty distribution. Honoré (1993) further shows that the finite mean or the tail condition can be removed with multiple spell data if there is no lagged duration dependence and if the frailty is time invariant.

For simplicity of exposition, we work with single spell uncensored observations. Following the multiplicative on hazard assumption in the identification literature, we assume the individual conditional hazard function to be of the form,

$$\lambda(t_i | x_i, U_i) = \lambda_0(t_i) \exp(x_i' \beta) v_i = \lambda_0(t_i) \exp(x_i' \beta + \xi U_i),$$

where v_i is the frailty term, further parametrized as $\exp(\xi U_i)$ with U_i follows distribution F with zero mean and unit variance. The mean of the frailty v_i is approximately 1 for ξ small. The baseline hazard function $\lambda_0(t_i)$ is known up to a finite number of parameters and the x_i 's denote the vector of covariates including an intercept term. We would like to evaluate the advisability of introducing the frailty term, that is, again, to test $H_0 : \xi = 0$ against $H_a : \xi \neq 0$.

The individual survival function and the unconditional density can be written as

$$\begin{aligned} S(t_i | x_i) &= \int e^{-\Lambda_0(t_i) \exp(x_i' \beta + \xi u_i)} dF(u_i) \\ f(t_i | x_i) &= \int \lambda_0(t_i) e^{x_i' \beta + \xi u_i} e^{-\Lambda_0(t_i) \exp(x_i' \beta + \xi u_i)} dF(u_i), \end{aligned}$$

It is sometimes called a mixed proportional hazard model in the literature because f is a mixture density and F is the mixing distribution. With different assumption on the baseline hazard, we have the following two examples.

3.2.1. Exponential Baseline Hazard. Assume that the baseline hazard follows exponential distribution with $\Lambda_0(t) = t$ and $\lambda_0(t) = 1$. The residual score for ξ is,

$$g(t_i, \beta) = (1 - 3t_i e^{x_i' \beta} + t_i^2 e^{2x_i' \beta}) + (1 - t_i e^{x_i' \beta}).$$

Replacing β by its MLEs, that $\sum_i (1 - t_i \exp(x_i' \hat{\beta})) x_i = 0$, the optimal $C(\alpha)$ test is

$$Z_n = \sum_i \left(1 - 3t_i \exp(x_i' \hat{\beta}) + t_i^2 \exp(2x_i' \hat{\beta}) \right) / \sqrt{4n}$$

This is identical to the test proposed in Kiefer (1984), which uses the approach in Cox (1983) and derives the score test based on whether the variance of the frailty term is zero in local approximation of the mixture density.

3.2.2. Weibull Baseline Hazard. With a slight modification, we can instead assume the baseline hazard as a Weibull model as done in Lancaster (1985) with unknown shape parameter α that $\lambda_0(t) = \alpha t^{\alpha-1}$. In this case, we have one more nuisance parameter α in addition to the coefficients β . Replacing the nuisance parameters by their respective MLEs, we find the $C(\alpha)$ test as

$$Z_n = \frac{\sum_i (1 - 3t_i^{\hat{\alpha}} \exp(x_i' \hat{\beta}) + t_i^{2\hat{\alpha}} \exp(2x_i' \hat{\beta}))}{\sqrt{4n - 4n/q}}$$

with $q = 1 + \psi'(2) - (\psi(2))^2$ in which $\psi'(z)$ is the trigamma function and $\psi(z)$ is the digamma function. (Details in the Appendix B). Lancaster (1985) obtains the same test statistic using, again, the approach of Cox (1983). We will discuss in more details in Section 4 the relation of the heterogeneity tests of Cox (1983) and Chesher (1984) to the $C(\alpha)$ test.

3.3. Joint test for location and scale heterogeneity in Gaussian panel data model.

In this example, we consider a two dimensional $C(\alpha)$ test for parameter heterogeneity in a Gaussian panel data model. The model is assumed to be

$$y_{it} = \mu_i + \sigma_i \epsilon_{it}$$

with $\mu_i = \mu_0 + \xi_1 U_{1i}$ and $\sigma_i^2 = \sigma_0^2 \exp(\xi_2 U_{2i}) \geq 0$. The random variables U_{ki} are i.i.d. with distribution F_k for $k = 1, 2$. Both U_1 and U_2 have zero mean and unit variance and are assumed to be independent.

The unconditional density of observing (y_{i1}, \dots, y_{iT}) is

$$f_i = \iint \left(\frac{1}{2\pi\sigma_0^2 \exp(\xi_2 u_{2i})} \right)^{T/2} \exp \left(- \sum_{t=1}^T \frac{(y_{it} - \mu_0 - \xi_1 u_{1i})^2}{2\sigma_0^2 \exp(\xi_2 u_{2i})} \right) dF_1(u_{1i}) dF_2(u_{2i})$$

The respective score for (ξ_1, ξ_2) and the nuisance parameters (μ_0, σ_0^2) are

$$\begin{aligned} v_{1i} &= \nabla_{\xi_1}^2 \log f_i |_{\xi_1=\xi_2=0} = \left(\frac{\bar{y}_i - \mu_0}{\sigma_0^2/T} \right)^2 - \frac{1}{\sigma_0^2/T} \\ v_{2i} &= \nabla_{\xi_2}^2 \log f_i |_{\xi_1=\xi_2=0} = \left(Z_i - \frac{T}{2} \right)^2 - Z_i \\ v_{3i} &= \nabla_{\mu_0} \log f_i |_{\xi_1=\xi_2=0} = \frac{\bar{y}_i - \mu_0}{\sigma_0^2/T} \\ v_{4i} &= \nabla_{\sigma_0^2} \log f_i |_{\xi_1=\xi_2=0} = (Z_i - \frac{T}{2})/\sigma_0^2 \end{aligned}$$

where \bar{y}_i is the sample mean defined as $\sum_{t=1}^T y_{it}/T$ and $2Z_i = \sum_{t=1}^T (y_{it} - \mu_0)^2/\sigma_0^2 \sim \chi_T^2$.

Replacing the nuisance parameters by their MLEs, the optimal $C(\alpha)$ test for $H_0 : \xi_1 = \xi_2 = 0$ against $H_a : \xi_i \neq 0$ for at least one i is:

$$T_n = (0 \vee t_{1n})^2 + (0 \vee t_{2n})^2$$

with

$$\begin{aligned} t_{1n} &= (2NT(T-1)/\hat{\sigma}_0^4)^{-1/2} \left(\sum_i \left(\frac{\bar{y}_i - \hat{\mu}_0}{\hat{\sigma}_0^2/T} \right)^2 - \frac{NT}{\hat{\sigma}_0^2} \right) \\ t_{2n} &= (NT(T/2+1))^{-1/2} \left(\sum_i \left(Z_i - \frac{T}{2} \right)^2 - \frac{NT}{2} \right) \end{aligned}$$

We reject H_0 for $T_n > c_\alpha$ where c_α is the $(1-\alpha)$ -quantile of $\frac{1}{4}\chi_0^2 + \frac{1}{2}\chi_1^2 + \frac{1}{4}\chi_2^2$.

Remark. The first component t_{1n} of the test statistics may be recognized as the test for individual effect in Gaussian panel data model proposed by Breusch and Pagan (1980). The second component t_{2n} is equivalent to a single parameter $C(\alpha)$ test for a Gamma model with heterogenous scale parameter. The factorization provided by the Gaussian model leads to simple asymptotics of the test statistics. Dependence between the random effects U_1 and U_2 would introduce more complicated weights for the χ^2 mixture as alluded to earlier discussion at the end of Section 2. (Computational details in the Appendix B.)

4. REPARAMETERIZATION AND CONNECTION TO THE INFORMATION MATRIX TEST

4.1. Reparameterization. A common strategy in prior literature to circumvent the irregularity, that the score function is degenerately zero, is to reparameterize the model. In fact, this is the advice given in the original Neyman's (1959) $C(\alpha)$ paper (Section 9, p. 225) and also in Cox and Hinkley (1974, p. 117-118). For the heterogeneity tests considered in

this paper in particular, Cox (1983) and Chesher (1984) adopt such a reparameterization by letting $\eta = \xi_0 + (\xi - \xi_0)^2$. Reconsidering the example in Section 2.2, without loss of generality, we set $\xi_0 = 0$ and have the density function as $p(x; \lambda_0 + \tau\sqrt{\eta}\mathbf{U}_i)$. Cox (1983) tests for heterogeneity of λ_i by testing $H_0 : \eta = 0$ against $H_1 : \eta > 0$. Chesher (1984) takes the same model assuming \mathbf{U}_i follows a symmetric location-scale distribution.

At first sight, reparameterization avoids the irregularity of having a degenerate score function. The first order derivative with respect to η , albeit an undefined $\frac{0}{0}$ function, can be evaluated by the l'Hôpital's rule. As long as $\mathbb{E}(\mathbf{U}^2)$ is non-zero, the score function is nonvanishing. The score test thus derived will be identical to the $C(\alpha)$ test using the original parameterization that $\lambda_i = \lambda_0 + \tau\xi\mathbf{U}_i$. However, the second order derivative for η is unbounded unless we impose an additional moment condition on \mathbf{U} , that $\mathbb{E}(\mathbf{U}^3) = 0$ (See derivation in the Appendix C). This condition is implicitly satisfied in Chesher (1984) because of the symmetry distribution assumption on \mathbf{U} . Moran (1973) also employed this zero third moment condition and remarked that it was hard to rationalize. One explanation for this extra condition is that the original, more natural specification on the random parameter $\lambda_i = \lambda_0 + \tau\xi\mathbf{U}_i$ with $\xi \in \mathbb{R}$ is not equivalent to the reparameterization $\lambda_i = \lambda_0 + \tau\sqrt{\eta}\mathbf{U}_i$ with $\eta \in \mathbb{R}_+$ unless \mathbf{U} has a symmetric distribution. As we have seen, the ξ parameterization has the advantage that no symmetry or higher moment conditions are necessary.

4.2. Connection to the Information Matrix test. Chesher (1984) also points out that White's (1982) Information Matrix (IM) test is a score test for unobserved heterogeneity. Since Chesher (1984) can be viewed as a reparameterized $C(\alpha)$ test, it is of interest to investigate the connection between the $C(\alpha)$ test for heterogeneity in general and the IM test.

Take again the example in Section 2.2, Y_1, \dots, Y_n are i.i.d. random variables each with density function $p(y; \lambda_i)$. The parameter λ_i is a random parameter and we assume it now takes a more general form $\lambda_i = \lambda_0 + \xi k(\lambda_0)\mathbf{U}_i$ to incorporate both additive and multiplicative specifications. For example, if $k(\lambda_0) = 1$, we have the additive form $\lambda_i = \lambda_0 + \xi\mathbf{U}_i$, while if $k(\lambda_0) = \lambda_0$, then the multiplicative form. The function $k(\lambda_0)$ thus allows flexible specification for the random parameter.

For simplicity and to fix ideas, we first assume λ_0 is known. Theorem 1 then implies the following expansion of the log-likelihood function, provided that $\xi_n = O(n^{-1/4})$,

$$l = \sum_i \log \int p(y_i; \lambda_i) dF(\mathbf{u}) = \sum_i \log p(y_i; \lambda_0) + \frac{1}{2} \xi_n^2 \mathbb{E}(\mathbf{U}_i^2) \sum_i k(\lambda_0)^2 \frac{\nabla_{\lambda}^2 p(y_i; \lambda_0)}{p(y_i; \lambda_0)} + O_p(1)$$

The first order derivative of l with respect to ξ_n is zero evaluated under $\xi_n = 0$, and the second-order score is

$$\frac{\partial^2}{\partial \xi_n^2} l|_{\xi_n=0} = \sum_i k(\lambda_0)^2 \frac{\nabla_{\lambda}^2 p(y_i; \lambda_0)}{p(y_i; \lambda_0)}.$$

If λ_0 is unknown, we find the corresponding score for λ_0 and take the projection step to get the $C(\alpha)$ test. This is very close to the approximation in Cox (1983) except we allow for a more flexible variance function for random parameter λ_i , as $\xi^2 \mathbb{E}(\mathbf{U}_i^2) k(\lambda_0)^2$. In a regression

model with covariates, λ_0 will then be a function of the covariates with coefficients β .

White's (1982) Information Matrix test under regression setting, on the other hand, is constructed based on the following moment conditions:

$$\mathbb{E} \left[\text{vech} \left(\nabla_{\beta}^2 \log p(y; \lambda_0(x_i, \beta)) + \nabla_{\beta} \log p(y; \lambda_0(x_i, \beta)) \nabla_{\beta}^{\top} \log p(y; \lambda_0(x_i, \beta)) \right) \right] = 0$$

where vech is the operator which stacks the elements in the lower triangular part of a symmetric matrix. By using the chain rule and focusing only on the moment condition for the intercept term β_0 , the IM test statistic uses the following sample analogue of the moment condition

$$\text{IM} = \sum_i \left[\frac{\nabla_{\lambda}^2 p(y; \lambda_0(x_i, \beta))}{p(y; \lambda_0(x_i, \beta))} (\nabla_{\beta_0} \lambda_0(x_i, \beta))^2 + \frac{\nabla_{\lambda} p(y; \lambda_0(x_i, \beta))}{p(y; \lambda_0(x_i, \beta))} \nabla_{\beta_0}^2 \lambda_0(x_i, \beta) \right]$$

For the $C(\alpha)$ test to be equivalent to the IM test, it is sufficient to have the following two identities:

$$\begin{aligned} C(\nabla_{\beta_0} \lambda_0(x_i, \beta))^2 &= k(\lambda_0(x_i, \beta))^2 \\ \sum_i \frac{\nabla_{\lambda} p(y; \lambda_0(x_i, \beta))}{p(y; \lambda_0(x_i, \beta))} \nabla_{\beta_0}^2 \lambda_0(x_i, \beta) &= 0 \end{aligned}$$

where C is a non-zero constant. We give several examples below as illustrations.

Example 4.1. *Normal regression with $Y_i \sim \mathcal{N}(\mu_i, 1)$, where $\mu_i = \mu_{0i} + \xi k(\mu_{0i}) \mathbf{U}_i$ and $\mu_{0i} = x_i' \beta$.*

Note that $\nabla_{\beta_0} \mu_{0i} = 1$ and $\nabla_{\beta_0}^2 \mu_{0i} = 0$, the IM test is equivalent to the $C(\alpha)$ test if $k(\mu_{0i}) = C \neq 0$ and all nuisance parameters be replaced by their MLEs. Any other functional form of $k(\mu_{0i})$ leads to a test statistic that is different from the IM test.

Example 4.2. *Poisson regression with $Y_i \sim \text{Poi}(\lambda_i)$, where $\lambda_i = \lambda_{0i} + \xi k(\lambda_{0i}) \mathbf{U}_i$ and $\lambda_{0i} = \exp(x_i' \beta)$.*

Now $\nabla_{\beta_0} \lambda_{0i} = \nabla_{\beta_0}^2 \lambda_{0i} = \lambda_{0i}$. If β 's are replaced by their MLEs, the second identify for equivalence holds because the normal equation for the MLE of β_0 gives

$$\sum_i \frac{\nabla_{\lambda} p(y; \lambda_{0i})}{p(y; \lambda_{0i})} \nabla_{\beta_0}^2 \lambda_{0i} = \sum_i \frac{\nabla_{\lambda} p(y; \lambda_{0i})}{p(y; \lambda_{0i})} \nabla_{\beta_0} \lambda_{0i} = 0$$

Therefore, the IM test is equivalent to the $C(\alpha)$ test if $k(\lambda_{0i}) = \lambda_{0i}$ which is satisfied for the multiplicative alternative $\lambda_i = \lambda_{0i}(1 + \xi \mathbf{U}_i)$. This specification is a first order linear approximation of the alternative form $\lambda_i = \lambda_{0i} \exp(\xi \mathbf{U}_i)$ for small ξ , which leads to the second moment test for the Poisson regression model as discussed in Section 3.1.1.

In summary, when the model contains covariates, the $C(\alpha)$ test is equivalent to the IM test only under a particular alternative specification, provided that the nuisance parameters are also replaced by their corresponding restricted MLEs. When the model does not contain covariates, IM test will be equivalent to the $C(\alpha)$ test because the function $k(\lambda_0)$ is no longer

individual specific and can be factored out as a constant from the score function. It will then be cancelled when we rescale the score by its standard deviation to form the $C(\alpha)$ test statistic. A similar conclusion is reached in several papers that observe that the score test for unobserved parameter heterogeneity is not always identical to the IM test. In particular, Cameron and Trivedi (1986) and Dean (1992) discusses the impact of different specification on the overdispersion test statistic for count data regression models.

5. CONCLUSION

We have shown that Neyman's $C(\alpha)$ test provides a unified approach to testing for neglected heterogeneity in parametric models. The irregularity encountered in these testing problems, that the score function is identically zero, can be circumvented by defining a second-order score function. Optimality of this new score function is established by formulating the problem in LeCam's LAN framework and examining the associated limit experiment. This framework provides neater regularity conditions in the irregular problem as compared to classical approach in Neyman (1959).

The $C(\alpha)$ test inherits the chief merit of the score test, computation is made easy under the null model. In contrast, the likelihood ratio test, in face of the generally unknown heterogeneity distribution F , is computationally challenging. We have also seen that the $C(\alpha)$ test has local power against a wide class of alternatives, that allows us to avoid strict parametric assumptions on F , relying instead on weaker moment conditions. A further advantage of the LeCam framework is that it enables us to dispense with symmetry and higher order moment conditions that have been employed in earlier work.

A straightforward generalization of the theorems in Section 2 would be to incorporate density functions that allow the first $(m - 1)$ logarithmic derivatives to vanish. Rotnitzky, Cox, Bottai, and Robins (2000) also discuss estimation problems in this general case under classical MLE type of conditions. In such cases, we can define the m^{th} order derivative of the log density as the score function and require the Pitman-type local alternative to be of order $n^{-1/2m}$. LeCam's DQM condition needs to be modified by raising the corresponding elements in the expansion to m^{th} power, as we did for $m = 2$ in Definition 1. It is curious to observe that only when m is an even integer is the test required to be one-sided. When m is odd, we can use reparameterization to transform the irregular problem back to a regular case, without imposing additional restrictions (i.e. symmetry of the distribution F).

A drawback of the $C(\alpha)$ test, as reflected in Neyman (1979), is that asymptotic optimality of the test is only established under local alternatives. The approximation of the power function, which is characterized by the asymptotic behavior of the test statistics under such alternatives, relies on n tending to infinity and the parameter ξ_n converging to the null value ξ_0 . The behavior of the power function for finite samples or fixed alternatives is largely unknown. It is also of interest to compare the power behavior of the $C(\alpha)$ test to the likelihood ratio test in these random coefficient models. We hope to pursue these investigations in future work.

REFERENCES

- BICKEL, P., C. KLAASSEN, Y. RITOV, AND J. WELLNER (1993): *Efficient and Adaptive estimation for semi parametric models*. Johns Hopkins University Press: Baltimore and London.
- BREUSCH, T., AND A. PAGAN (1980): "The Lagrange Multiplier test and its applications to model specification in Econometrics," *Review of Economic Studies*, 47, 239–253.
- BÜLHER, W., AND P. PURI (1966): "On optimal asymptotic tests of composite hypotheses with several constraints," *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 5, 71–88.
- CAMERON, A., AND P. TRIVEDI (1986): "Econometric models based on count data: comparisons and applications of some estimators and tests," *Journal of Applied Econometrics*, 1, 29–53.
- CHEN, H., J. CHEN, AND J. KALBFLEISCH (2001): "A modified likelihood ratio test for homogeneity in finite mixture models," *Journal of the Royal Statistical Society, Series B*, 63(1), 19–29.
- CHERNOFF, H. (1954): "On the distribution of the likelihood ratio," *The Annals of Mathematical Statistics*, 25(3), 573–578.
- CHESHER, A. (1984): "Testing for neglected heterogeneity," *Econometrica*, 52(4), 865–872.
- COLLINGS, B., AND B. MARGOLIN (1985): "Testing Goodness of Fit for the Poisson assumption when observations are not identically distributed," *Journal of the American Statistical Association*, 80, 411–418.
- COX, D. (1983): "Some remarks on overdispersion," *Biometrika*, 70(1), 269–274.
- COX, D., AND D. HINKLEY (1974): *Theoretical Statistics*. Chapman and Hall: London.
- CRAMÉR, H. (1946): *Mathematical Methods of Statistics*. Princeton University Press: Princeton, New Jersey.
- DEAN, C. (1992): "Testing for Overdispersion in Poisson and Binomial regression models," *Journal of the American Statistical Association*, 87, 451–457.
- DEAN, C., AND J. LAWLESS (1989): "Tests for detecting over dispersion in Poisson regression models," *Journal of the American Statistical Association*, 84, 467–472.
- ELBERS, C., AND G. RIDDER (1982): "True and Spurious duration dependence: The identifiability of the Proportional Hazard model," *The Review of Economic Studies*, 49(3), 403–409.
- FISHER, R. (1950): "The significance of deviations from expectation in a Poisson series," *Biometrika*, 6, 17–24.
- GOURIÉROUX, C., A. HOLLY, AND A. MONFORT (1982): "Likelihood ratio test, wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters," *Econometrica*, 50(1), 63–80.
- HALLEN, M., AND C. LEY (2012): "Skew-Symmetric Distributions and Fisher Information: The Double Sin of the Skew-Normal," arXiv: 1209.4177[math.ST].
- HECKMAN, J., AND B. SINGER (1982): "The identification problem in econometric models for duration data," in *Advances in Econometrics*, ed. by W. Hildenbrand. Cambridge University Press.
- (1984): "The identifiability of the Proportional hazard model," *The Review of Economic Studies*, 51(2), 231–241.
- HILLIER, G. (1986): "Joint tests for zero restrictions on nonnegative regression coefficients," *Biometrika*, 73(3), 657–669.
- HONORÉ, B. (1993): "Identification results for duration models with multiple spells," *Review of Economic Studies*, 60(1), 21–246.
- KIEFER, N. (1984): "A Simple test for heterogeneity in exponential models of duration," *Journal of Labor Economics*, 2(4), 539–549.
- LANCASTER, T. (1985): "Generalized residuals and heterogenous duration models with applications to the Weibull model," *Journal of Econometrics*, 28, 155–169.
- LANCASTER, T., AND S. NICKELL (1980): "The analysis of re-employment probabilities for unemployed," *Journal of the Royal Statistical Society, Series A*, pp. 141–165.
- LECAM, L. (1972): "Limits of Experiments," in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, pp. 245–261. University of California Press: Berkeley and Los Angeles.
- LEE, L. (1986): "Specification test for Poisson regression models," *International Economic Review*, 27(3), 689–706.
- LEE, L., AND A. CHESHER (1986): "Specification testing when score test statistics are identically zero," *Journal of Econometrics*, 31, 121–149.

- MORAN, P. (1973): "Asymptotic properties of homogeneity tests," *Biometrika*, 60(1), 79–85.
- NEYMAN, J. (1959): "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics, the Harald Cramer Volume*, ed. by U. Grenander. Wiley: New York.
- (1979): " $C(\alpha)$ tests and their use," *Sankhyā: The Indian Journal of Statistics*, 41, 1–21.
- NEYMAN, J., AND E. SCOTT (1966): "On the use of $C(\alpha)$ tests of composite hypotheses," *Bull. Inst. Int. Statist.*, 41(1), 477–497.
- POLLARD, D. (1997): "Another look at differentiability in quadratic mean," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, ed. by D. Pollard, E. Torgersen, and G. Yang. Springer-Verlag: New York.
- RAO, C. (1948): "Large-Sample Test of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," *Proceedings of the Cambridge Philosophical Society*, 44, 50–57.
- (1973): *Linear Statistical Inference and its Applications*. Wiley.
- ROTNITZKY, A., D. COX, M. BOTTAI, AND J. ROBINS (2000): "Likelihood-based inference with singular information matrix," *Bernoulli*, 6(2), 243–284.
- SELF, S., AND K.-Y. LIANG (1987): "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions," *Journal of the American Statistical Association*, 82(398), 605–610.
- VAN DER VAART, A. (1991): "An asymptotic representation theorem," *International Statistical Review*, 59(1), 97–121.
- (1998): *Asymptotic Statistics*. Cambridge University Press.
- VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes - Springer Series in Statistics*. Springer: New York.
- VAUPEL, J., K. MANTON, AND W. STALLARD (1979): "The impact of heterogeneity in individual frailty on the dynamics of mortality," *Demography*, 16(3), 439–454.
- WHITE, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica*, 50, 1–25.

APPENDIX A. PROOF OF THEOREMS

Before proceeding to the proof for **Theorem 1**, we first prove the following lemma as an adaption to Pollard (1997, Lemma 1). Denote $f_n = \sqrt{p(x_i; \xi_n, \theta_n)}$ and $f_0 = \sqrt{p(x_i; \xi_0, \theta)}$. Let v_ξ and v_θ be shorthand for $v_\xi(x_i)$ and $v_\theta(x_i)$. Let $\|\cdot\|$ be $\mathcal{L}_2(\mu)$ -norm and $\langle \cdot, \cdot \rangle$ be the inner product. If it contains a vector, then it is defined as the vector of inner product for each elements. Further, let $r_n(x_i, \xi_n, \theta_n) = f_n - f_0 - h_n^\top v(x_i)$ and denote $R_i = r_n(x_i, \xi_n, \theta_n)/f_0$.

Lemma 1. Under Assumption 1 and the modified DQM condition, we have the following:

- (1) $\sum_i R_i^2 = o_p(1)$
- (2) $\mathbb{E}(v(X)/f_0) = 0$
- (3) $2 \sum_i R_i = -\frac{1}{4} t^\top J t + o_p(1)$
- (4) $n^{-1/2} \sum_i R_i v_\xi / f_0 = o_p(1)$, $n^{-1/2} \sum_i R_i v_\theta / f_0 = o_p(1)$
- (5) $\max_{1 \leq i \leq n} |R_i| = o_p(1)$
- (6) $\max_{1 \leq i \leq n} |\frac{2}{\sqrt{n}} \frac{v_\xi}{f_0}| = o_p(1)$, $\max_{1 \leq i \leq n} |\frac{2}{\sqrt{n}} \frac{v_\theta}{f_0}| = o_p(1)$

Proof of (1). Under the modified DQM condition, the Markov inequality yields,

$$\begin{aligned} \mathbb{P}(\sum_i R_i^2 > \epsilon) &\leq \epsilon^{-2} n \mathbb{E}(R_i^2) \\ &= \epsilon^{-2} n \int r_n^2(x; \xi_n, \theta_n) d\mu(x) \rightarrow 0. \end{aligned}$$

Proof of (2) and (3). Since both f_n and f_0 are objects with $\mathcal{L}_2(\mu)$ -norm 1

$$\begin{aligned} 0 &= \|f_n\|_{\mu,2}^2 - \|f_0\|_{\mu,2}^2 \\ &= (\xi_n - \xi_0)^4 \|\mathbf{v}_\xi\|_{\mu,2}^2 + (\theta_n - \theta)^\top \|\mathbf{v}_\theta\|_{\mu,2}^2 (\theta_n - \theta) + \|\mathbf{r}_n\|_{\mu,2}^2 + 2 \langle (\theta_n - \theta)^\top \mathbf{v}_\theta, \mathbf{r}_n \rangle \\ &\quad + 2(\xi_n - \xi_0)^2 (\theta_n - \theta)^\top \langle \mathbf{v}_\theta, \mathbf{v}_\xi \rangle + 2(\xi_n - \xi_0)^2 \langle \mathbf{v}_\xi, \mathbf{r}_n \rangle + 2(\xi_n - \xi_0)^2 \langle f_0, \mathbf{v}_\xi \rangle \\ &\quad + 2(\theta_n - \theta)^\top \langle f_0, \mathbf{v}_\theta \rangle + 2 \langle f_0, \mathbf{r}_n \rangle \end{aligned}$$

Note that by Cauchy-Schwarz inequality and the fact that both \mathbf{v}_ξ and \mathbf{v}_θ are square integrable with respect to measure μ by assumption, $\langle \mathbf{v}_\xi, \mathbf{r}_n \rangle = o(1/\sqrt{n})$ and $\langle \mathbf{v}_\theta, \mathbf{r}_n \rangle = o(1/\sqrt{n})$. Therefore, the fourth to sixth terms are all of order $o(1/n)$. The seventh and eighth term are both of order $O(1/\sqrt{n})$, so in order for the identity to hold, we must have

$$\langle f_0, \mathbf{v}_\xi \rangle = \langle f_0, \mathbf{v}_\theta \rangle = 0$$

This proves (2) since $0 = \langle f_0, \mathbf{v}_\xi \rangle = \mathbb{E}(\mathbf{v}_\xi(X)/f_0)$. Similar argument shows $\mathbb{E}(\mathbf{v}_\theta(X)/f_0) = 0$. Hence,

$$\begin{aligned} 2 \langle f_0, \mathbf{r}_n \rangle &= -(\xi_n - \xi_0)^4 \|\mathbf{v}_\xi\|_{\mu,2}^2 - (\theta_n - \theta)^\top \|\mathbf{v}_\theta\|_{\mu,2}^2 (\theta_n - \theta) \\ &\quad - 2(\xi_n - \xi_0)^2 (\theta_n - \theta)^\top \langle \mathbf{v}_\xi, \mathbf{v}_\theta \rangle + o(1/n) \\ &= -\frac{1}{4n} \mathbf{t}^\top \mathbf{J} \mathbf{t} + o(1/n) \end{aligned}$$

with $\mathbf{t}^\top = (\delta_1^2, \delta_2^\top)$.

Since $\mathbb{V}(2 \sum_i \mathbf{R}_i)$ is bounded above by $4 \sum_i \mathbb{E}(\mathbf{R}_i^2)$, which goes to 0 from (1), we have

$$\begin{aligned} 2 \sum_i \mathbf{R}_i &= 2n\mathbb{E}(\mathbf{R}_1) + o_P(1) \\ &= 2n \langle f_0, \mathbf{r}_n \rangle + o_P(1) \\ &= 2n \left(-\frac{1}{8n} \mathbf{t}^\top \mathbf{J} \mathbf{t} + o(1/n) \right) + o_P(1) \\ &= -\frac{1}{4} \mathbf{t}^\top \mathbf{J} \mathbf{t} + o_P(1) \end{aligned}$$

Proof of (4). By Hölder's inequality,

$$\sum_i \mathbf{R}_i \frac{2}{\sqrt{n}} \frac{\mathbf{v}_\xi}{f_0} \leq \sqrt{\sum_i \mathbf{R}_i^2 \sum_i \left(\frac{2}{\sqrt{n}} \frac{\mathbf{v}_\xi}{f_0} \right)^2} = o_P(1) O_P(1) = o_P(1)$$

Similar argument admits the second result.

Proof of (5).

$$\mathbb{P}(\max_{1 \leq i \leq n} |\mathbf{R}_i| > \epsilon) \leq n\mathbb{P}(|\mathbf{R}_1|^2 > \epsilon^2) \leq \epsilon^{-2} n\mathbb{E}(\mathbf{R}_1^2) \rightarrow 0$$

Proof of (6).

$$\begin{aligned} \mathbb{P}(\max_{1 \leq i \leq n} |2\mathbf{v}_\xi/f_0| > \epsilon\sqrt{n}) &\leq n\mathbb{P}(|2\mathbf{v}_\xi/f_0| > \epsilon\sqrt{n}) \\ &\leq \epsilon^{-2} \mathbb{E}((2\mathbf{v}_\xi(X_1)/f_0)^2) \mathbb{I}_{[|2\mathbf{v}_\xi/f_0| > \epsilon\sqrt{n}]} \rightarrow 0 \end{aligned}$$

Similar argument admits the second statement. ■

Proof of Theorem 1 We consider $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and $\theta_n = \theta + \delta_2 n^{-1/2}$ throughout the proof. Under **Assumption 1**, we have the following Taylor expansion:

$$f_n = f_0 + (\xi_n - \xi_0)^2 \mathbf{v}_\xi + (\theta_n - \theta)^\top \mathbf{v}_\theta + r_n(x_i; \xi_n, \theta_n).$$

Denoting $w_i = 2(f_n/f_0 - 1)$, we have

$$w_i = 2(\xi_n - \xi_0)^2 \frac{v_\xi}{f_0} + 2(\theta_n - \theta)^\top \frac{v_\theta}{f_0} + 2R_i.$$

To show that under the modified DQM condition, the log-likelihood ratio admits a quadratic approximation, we use results in **Lemma 1**.

The log-likelihood ratio can be represented as

$$\begin{aligned} \Lambda_n &= \sum_i \log \frac{p(x_i; \xi_n, \theta_n)}{p(x_i; \xi_0, \theta)} = \sum_i 2 \log \frac{f_n}{f_0} = \sum_i 2 \log(1 + w_i/2) \\ &= \sum_i w_i - \frac{1}{4} \sum_i w_i^2 + \frac{1}{2} \sum_i w_i^2 \beta(w_i) \end{aligned}$$

with $\beta(x) \rightarrow 0$ as $x \rightarrow 0$.

Using (3) in **Lemma 1** and with $S_n = (S_{\xi,n}, S_{\theta,n}^\top)^\top$ and J defined in **Theorem 1**, we have

$$\sum_i w_i = 2 \frac{\delta_1^2}{\sqrt{n}} \sum_i \frac{v_\xi}{f_0} + 2 \frac{\delta_2^\top}{\sqrt{n}} \sum_i \frac{v_\theta}{f_0} + 2 \sum_i R_i = t^\top S_n - \frac{1}{4} t^\top J t + o_p(1)$$

Using (1) and (4) in **Lemma 1**, we have

$$\begin{aligned} \sum_i w_i^2 &= \sum_i \left(\frac{2\delta_1^2}{\sqrt{n}} \frac{v_\xi}{f_0} + \frac{2\delta_2^\top}{\sqrt{n}} \frac{v_\theta}{f_0} + 2R_i \right)^2 \\ &= t^\top J t + o_p(1) + 4 \sum_i R_i^2 + 4 \sum_i R_i \left(\frac{2\delta_1^2}{\sqrt{n}} \frac{v_\xi}{f_0} + \frac{2\delta_2^\top}{\sqrt{n}} \frac{v_\theta}{f_0} \right) \\ &= t^\top J t + o_p(1) \end{aligned}$$

Lastly, we need to show that $\sum_i w_i^2 \beta(w_i) = o_p(1)$. First note that using (5) and (6) in **Lemma 1**, we have

$$\begin{aligned} \mathbb{P} \left(\max_{1 \leq i \leq n} |w_i| > \epsilon \right) &\leq \delta_1^2 \mathbb{P} \left(\max_{1 \leq i \leq n} \left| \frac{2}{\sqrt{n}} \frac{v_\xi}{f_0} \right| > \epsilon \right) + \delta_2^\top \mathbb{P} \left(\max_{1 \leq i \leq n} \left| \frac{2}{\sqrt{n}} \frac{v_\theta}{f_0} \right| > \epsilon \right) \\ &\quad + 2 \mathbb{P} \left(\max_{1 \leq i \leq n} |R_i| > \epsilon \right) \rightarrow 0 \end{aligned}$$

Since when $w_i \rightarrow 0$, $\beta(w_i) \rightarrow 0$, we have $\max_{1 \leq i \leq n} |\beta(w_i)| = o_p(1)$. By Hölder's inequality,

$$\sum_i w_i^2 \beta(w_i) \leq \max_{1 \leq i \leq n} |\beta(w_i)| \sum_i w_i^2 = o_p(1) O_p(1) = o_p(1).$$

Therefore, the log-likelihood ratio is approximated by

$$\begin{aligned} \Lambda_n &= \sum_i w_i - \frac{1}{4} \sum_i w_i^2 + \frac{1}{2} \sum_i w_i^2 \beta(w_i) \\ &= t^\top S_n - \frac{1}{4} t^\top J t - \frac{1}{4} t^\top J t + o_p(1) \\ &= t^\top S_n - \frac{1}{2} t^\top J t + o_p(1) \end{aligned}$$

■

Proof of Corollary 1 Since S_n is a normed iid sum, by the central limit theorem,

$$S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J)$$

The zero asymptotic mean of S_n is provided by (2) in **Lemma 1**, then the asymptotic variance for S_n is J as defined in Theorem 1.

The quadratic approximation for Λ_n established in **Theorem 1** together with the joint normality of S_n leads to the LAN property of the sequence of model P_{n,ξ_n,θ_n} . Furthermore, we have

$$\Lambda_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}\left(-\frac{1}{2}t^\top J t, t^\top J t\right).$$

By LeCam's first lemma (see e.g. van der Vaart (1998, Lemma 6.4)), P_{n,ξ_n,θ_n} and $P_{n,\xi_0,\theta}$ are mutually contiguous. \blacksquare

Proof of Theorem 2 The sequence of experiments \mathcal{E}_n converges to a shifted Gaussian $\mathcal{N}(t, J^{-1})$ as a result of Theorem 9.4 in van der Vaart (1998). The log-likelihood ratio process of observing one sample from $\mathcal{N}(t, J^{-1})$ is

$$\log \frac{d\mathcal{N}(t, J^{-1})}{d\mathcal{N}(0, J^{-1})}(Y) = t^\top J Y - \frac{1}{2}t^\top J t$$

It suffices to show that $J^{-1}S_n$ converges to the distribution of Y under the null. **Corollary 1** establishes $S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J)$, we thus have $J^{-1}S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J^{-1})$.

The optimal test statistic for $H_0 : \delta_1 = 0$ against $H_a : \delta_1 \neq 0$ in the limit experiment is the first element in Y . The sequence of test statistics from the original experiment \mathcal{E}_n that matches with the first element in Y is the $C(\alpha)$ statistic,

$$Z_n = (J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{-1/2}(S_{\xi,n} - J_{\xi\theta}J_{\theta\theta}^{-1}S_{\theta,n}).$$

Notice the rescaling in Z_n is needed to obtain a unit asymptotic variance for the test statistic. \blacksquare

Proof of Corollary 2 Since ξ is a scalar and $S_n \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, J)$ under H_0 , it is immediate that the asymptotic null distribution for Z_n is $\mathcal{N}(0, 1)$.

We can now use LeCam's third lemma (see e.g. van der Vaart (1998, Example 6.7)) to derive the asymptotic distribution for Z_n under local alternatives. We are interested in the local alternative that $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and nuisance parameter θ is left unspecified as in the null, hence we set $\delta_2 = 0$ in the log-likelihood ratio expansion. Under H_0 ,

$$(Z_n, \Lambda_n) \overset{P_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\delta_1^4 J_{\xi\xi} \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \delta_1^4 J_{\xi\xi} \end{pmatrix}\right)$$

with $\sigma_{12} = \text{Cov}(Z_n, \Lambda_n) = \delta_1^2 (J_{\xi\xi} - J_{\xi\theta}J_{\theta\theta}^{-1}J_{\theta\xi})^{1/2}$. With $\delta_2 = 0$, **Corollary 1** implies that $P_{n,\xi_n,\theta}$ are mutually contiguous to $P_{n,\xi_0,\theta}$, then LeCam's third lemma implies,

$$Z_n \overset{P_{n,\xi_n,\theta}}{\rightsquigarrow} \mathcal{N}(\sigma_{12}, 1).$$

■

Proof of Theorem 3 Define the class of functions:

$$\mathcal{F}_n := \left\{ x \mapsto (g(x, \theta) - g(x, \eta)) \mid \|\theta - \eta\| \leq \delta_n \right\}.$$

If $\hat{\theta}$ is a \sqrt{n} -consistent estimator of θ , and $\delta_n = O(n^{-k})$ with $k < 1/2$, we obtain that with probability tending to one

$$\left| Z_n(\hat{\theta}) - Z_n(\theta) \right| \leq \sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)|$$

where $\mathbb{G}_n(f) := n^{-1/2} \sum_i (f(X_i) - \mathbb{E}f(X_i))$ denotes the empirical process indexed by \mathcal{F}_n . Proving $Z_n(\hat{\theta}) - Z_n(\theta) = o_P(1)$ thus amounts to establishing asymptotic equicontinuity of the process \mathbb{G}_n with respect to the Euclidean norm.

Let the parameter space near true θ , $\mathcal{U}_{\delta_n}(\theta)$, be covered by balls with radius $\epsilon^{1/\gamma}$, the number of balls can be upper bounded by $C_1 \epsilon^{-p/\gamma}$ with C_1 as a constant that does not depend on n and p being the dimension of the nuisance parameter space. Then for $\forall \eta \in \mathcal{U}_{\delta_n}(\theta)$, $\exists N_\eta$, such that

$$\|\eta - \eta_{N_\eta}\| \leq \epsilon^{1/\gamma}$$

The condition on g in **Assumption 2** implies

$$|g(x, \eta) - g(x, \eta_{N_\eta})| \leq \|\eta - \eta_{N_\eta}\|^\gamma H(x) \leq \epsilon H(x)$$

It follows that the bracketing number, $N_{[\cdot]}(\epsilon \|H\|_2, \mathcal{F}_n, \mathcal{L}_2(P_{n, \xi_n, \theta}))$ is bounded from above by $C_2 \epsilon^{-p/\gamma}$.

Furthermore, the assumption also implies that for $f \in \mathcal{F}_n$, $\|f\|_{P_{n,2}} \leq \delta_n^\gamma \|H\|_{P_{n,2}}$ with $\mathcal{L}_2(P_{n, \xi_n, \theta})$ -norm. We can now apply Theorem 2.14.2 in van der Vaart and Wellner (1996) and get

$$\mathbb{E}_{P_{n, \xi_n, \theta}} \left(\sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)| \right) \leq J_{[\cdot]}(\delta_n^\gamma, \mathcal{F}_n, \mathcal{L}_2(P_{n, \xi_n, \theta})) \|H\|_{P_{n,2}} + \sqrt{n} \mathbb{E}_{P_{n, \xi_n, \theta}} [H(X) I\{H(X) > \sqrt{n} \alpha(\delta_n^\gamma)\}]$$

where the bracketing integral is defined as

$$J_{[\cdot]}(\delta_n^\gamma, \mathcal{F}_n, \mathcal{L}_2(P_{n, \xi_n, \theta})) = \int_0^{\delta_n^\gamma} \sqrt{1 + \log N_{[\cdot]}(\epsilon \|H\|_{P_{n,2}}, \mathcal{F}_n, \mathcal{L}_2(P_{n, \xi_n, \theta}))} d\epsilon$$

and

$$\alpha(\delta_n^\gamma) = \delta_n^\gamma \|H\|_2 / \sqrt{1 + \log N_{[\cdot]}(\delta_n^\gamma \|H\|_{P_{n,2}}, \mathcal{F}_n, \mathcal{L}_2(P_{n, \xi_n, \theta}))}.$$

Provided that $\delta_n \rightarrow 0$, we have for n large enough,

$$J_{[\cdot]}(\delta_n^\gamma, \mathcal{F}_n, \mathcal{L}_2(P_{n, \xi_n, \theta})) \leq \int_0^{\delta_n^\gamma} 1 + \log(C_2 \epsilon^{-p/\gamma}) d\epsilon \rightarrow 0$$

Since $H(x)$ is square integrable for all n by **Assumption 2**, the first term goes to zero.

The upper bound for the bracketing number also yields a lower bound for $\alpha(\delta_n^\gamma)$ that is for δ_n sufficiently small,

$$\alpha(\delta_n^\gamma) \geq \frac{\delta_n^\gamma \|H\|_{P_{n,2}}}{\sqrt{1 + \log(C_2 \delta_n^{-p})}} := k_n \rightarrow 0$$

As long as k_n converges to zero slower than c_n , **Assumption 2** ensures that the second term also tends to zero.

The last step is to check that $\sup_{f \in \mathcal{F}_n} \frac{1}{\sqrt{n}} \sum_i \mathbb{E}_{P_{n,\xi_n,\theta}}(f(X_i)) = o(1)$ so that $\sup_{f \in \mathcal{F}_n} |\mathbb{G}_n(f)|$ is the correct upper bound. This is trivially true under the null, where $\xi_n = \xi_0$ for all $n \in \mathbb{N}$, since $\mathbb{E}_{P_{n,\xi_0,\theta}}(g(X_i, \theta)) = \mathbb{E}_{P_{n,\xi_0,\theta}}(g(X_i, \hat{\theta})) = 0$. Under local alternatives with $\xi_n = \xi_0 + \delta_1 n^{-1/4}$ and given the i.i.d. assumption on the sample, it suffices to show that

$$\sup_{\|\eta - \theta\| \leq \delta_n} \sqrt{n} \int (g(x, \eta) - g(x, \theta)) p(x; \xi_n, \theta) dx = o(1)$$

Denote $p_n = p(x; \xi_n, \theta)$ and $p_0 = p(x; \xi_0, \theta)$, we have the following expansion

$$\begin{aligned} \sqrt{n} \int (g(x, \eta) - g(x, \theta)) p_n dx &= \sqrt{n} \int \left((g(x, \eta) - g(x, \theta)) (\sqrt{p_0} + (\xi_n - \xi_0)^2 v_\xi(x) + r_n) \right) \sqrt{p_n} dx \\ &= \sqrt{n} \int (g(x, \eta) - g(x, \theta)) \sqrt{p_0} \sqrt{p_n} dx \\ &\quad + \sqrt{n} (\xi_n - \xi_0)^2 \int (g(x, \eta) - g(x, \theta)) \sqrt{p_n} v_\xi(x) dx \\ &\quad + \sqrt{n} \int (g(x, \eta) - g(x, \theta)) \sqrt{p_n} r_n dx \end{aligned}$$

The last two terms are $o(1)$ uniformly over η for $\|\eta - \theta\| \leq \delta_n$ due to the DQM condition in **Definition 1** and assumption on g in **Assumption 2**. Since Cauchy-Schwarz inequality implies that with respect to $\mathcal{L}_2(\mu)$ -norm,

$$\begin{aligned} \left| \int (g(x, \eta) - g(x, \theta)) \sqrt{p_n} v_\xi(x) dx \right| &\leq \|(g(x, \eta) - g(x, \theta)) \sqrt{p_0}\|_{\mu,2} \|v_\xi\|_{\mu,2} \\ &\leq \|\eta - \theta\|^\gamma \|H\|_{P_{n,2}} \|v_\xi\|_{\mu,2} = o(1). \end{aligned}$$

Similarly,

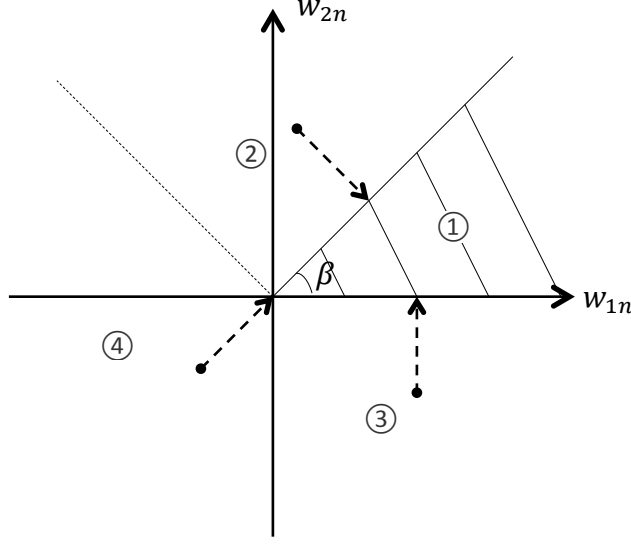
$$|\sqrt{n} \int (g(x, \eta) - g(x, \theta)) \sqrt{p_n} r_n dx| \leq \|(g(x, \eta) - g(x, \theta)) \sqrt{p_0}\|_{\mu,2} \sqrt{n} \|r_n\|_{\mu,2} = o(1).$$

The first term is also $o(1)$ by expanding $\sqrt{p_n}$ again and applying Cauchy-Schwarz inequality in a similar fashion. \blacksquare

Proof of Theorem 4 As in the proof of **Theorem 2**, the limit of the sequence v_n is a shifted Gaussian experiment $Y \sim \mathcal{N}(t, J^{-1})$ but now with $t^\top = (\delta_1^2, \delta_2^2, \delta_3^\top)$. An equivalent limit experiment observes $X \sim \mathcal{N}(t^\top J, J)$ with $X = JY$, because the likelihood ratio process of $\frac{dN(t, J^{-1})}{dN(0, J^{-1})}(Y)$ is identical to that of $\frac{dN(t^\top J, J)}{dN(0, J)}(X)$.

To be more explicit, denoting the first two elements of X to be X_ξ , and the rest to be X_θ , we have under the alternative,

$$\begin{pmatrix} X_\xi \\ X_\theta \end{pmatrix} \stackrel{\mathcal{D}}{=} \mathcal{N} \left((t_\xi, t_\theta) \begin{pmatrix} J_{\xi\xi} & J_{\xi\theta} \\ J_{\theta\xi} & J_{\theta\theta} \end{pmatrix}, J \right)$$



with $\mathbf{t}_\xi = (\delta_1^2, \delta_2^2)$ and $\mathbf{t}_\theta = \delta_3^\top$.

To focus on testing for zero restrictions on \mathbf{t}_ξ , we find the conditional distribution of \mathbf{X}_ξ on \mathbf{X}_θ to be

$$\tilde{\mathbf{X}}_\xi = \mathbf{X}_\xi - \mathbf{J}_{\xi\theta} \mathbf{J}_{\theta\theta}^{-1} \mathbf{X}_\theta \stackrel{\mathcal{D}}{=} \mathcal{N}(\mathbf{t}_\xi (\mathbf{J}_{\xi\xi} - \mathbf{J}_{\xi\theta} \mathbf{J}_{\theta\theta}^{-1} \mathbf{J}_{\theta\xi}), \mathbf{J}_{\xi\xi} - \mathbf{J}_{\xi\theta} \mathbf{J}_{\theta\theta}^{-1} \mathbf{J}_{\theta\xi}).$$

The matched statistic from the original experiment is then

$$\tilde{\mathbf{S}}_{\xi,n} = \mathbf{S}_{\xi,n} - \mathbf{J}_{\xi\theta} \mathbf{J}_{\theta\theta}^{-1} \mathbf{S}_{\theta,n}$$

Under H_0 , $\tilde{\mathbf{S}}_{\xi,n}$ follows $\mathcal{N}(0, \Sigma)$ with $\Sigma = \mathbf{J}_{\xi\xi} - \mathbf{J}_{\xi\theta} \mathbf{J}_{\theta\theta}^{-1} \mathbf{J}_{\theta\xi}$, and under local alternative, its asymptotic distribution is $\mathcal{N}(\mathbf{t}_\xi \Sigma, \Sigma)$.

Let the Cholesky decomposition of Σ be such that $\Lambda \Lambda^\top = \Sigma$, then $\Lambda^{-1} \tilde{\mathbf{S}}_{\xi,n} \stackrel{\mathbb{P}_{n,\xi_0,\theta}}{\rightsquigarrow} \mathcal{N}(0, \mathbf{I})$ and $\Lambda^{-1} \tilde{\mathbf{S}}_{\xi,n} \stackrel{\mathbb{P}_{n,\xi_n,\theta_n}}{\rightsquigarrow} \mathcal{N}(\mathbf{t}_\xi \Lambda, \mathbf{I})$. Since $\mathbf{t}_\xi = (\delta_1^2, \delta_2^2) \in \mathbb{R}_+^2$ and

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} := \mathbf{t}_\xi \Lambda = \begin{pmatrix} \delta_1^2 \sqrt{v_1} + \rho \delta_2^2 \sqrt{v_2} \\ \sqrt{v_2} \sqrt{1 - \rho^2} \delta_2^2 \end{pmatrix}$$

The feasible parameter set is therefore the convex cone defined as,

$$\left\{ (\eta_1, \eta_2) \mid \eta_2 \geq 0, \eta_1 - \frac{\rho}{\sqrt{1 - \rho^2}} \eta_2 \geq 0 \right\}.$$

For test statistic taking a value that falls outside of the feasible set, it needs to be projected onto the set. This yields the following four cases as illustrated in the figure.

Case 1: When the value of the test statistic \mathbf{w}_n falls into shaded area ①, the optimal test statistics is the sum of squares of the elements of \mathbf{w}_n :

$$T_n = w_{1n}^2 + w_{2n}^2 \sim \chi_2^2$$

Case 2: When the test statistic falls into area ②, we need to project w_n onto the convex cone ①, which gives a point with coordinates $(\rho^2 w_{1n} + \rho\sqrt{1-\rho^2}w_{2n}, \rho\sqrt{1-\rho^2}w_{1n} + (1-\rho^2)w_{2n})$. The $C(\alpha)$ test statistic is hence:

$$\begin{aligned} T_n &= (\rho^2 w_{1n} + \rho\sqrt{1-\rho^2}w_{2n})^2 + (\rho\sqrt{1-\rho^2}w_{1n} + (1-\rho^2)w_{2n})^2 \\ &= (\rho w_{1n} + \sqrt{1-\rho^2}w_{2n})^2 \sim \chi_1^2 \end{aligned}$$

Case 3: When the test statistic w_n falls in area ③, projecting onto the region ① yields $(w_{1n}, 0)$ and thus,

$$T_n = w_{1n}^2 \sim \chi_1^2$$

Case 4: Lastly, when w_n falls into region ④, projecting onto region ① yields $(0, 0)$ and hence,

$$T_n = 0 \sim \chi_0^2$$

The asymptotic distribution of the $C(\alpha)$ test statistics is a mixture of χ^2 's, for which the weights are characterized by the probability of falling into different regions. The angle β spanned by the shaded area ① as marked in the figure is $\beta = \cos^{-1}(\rho)$, hence the probability of falling into region ① is $\frac{\beta}{2\pi}$. The probability of falling into ② and ③ is $\frac{1}{2}$, leaves the probability of falling into ④ as $(\frac{1}{2} - \frac{\beta}{2\pi})$. ■

APPENDIX B. COMPUTATIONAL DETAILS IN EXAMPLES

B.1. Cox Proportional Hazard model with frailty. The second-order score for ξ takes the form,

$$\begin{aligned} &\nabla_{\xi}^2 \log f(t_i | x_i) |_{\xi=0} \\ &= \frac{\int \lambda_0(t_i) e^{x_i' \beta + \xi u_i} e^{-\Lambda_0(t_i)} e^{x_i' \beta + \xi u_i} [(1 - \Lambda_0(t_i) e^{x_i' \beta + \xi u_i})^2 u_i^2 - u_i^2 \Lambda_0(t_i) e^{x_i' \beta + \xi u_i}] dF(u_i) |_{\xi=0}}{\int \lambda_0(t_i) e^{x_i' \beta + \xi u_i} e^{-\Lambda_0(t_i)} e^{x_i' \beta + \xi u_i} dF(u_i)} \\ &= \mathbb{E}(U_i^2) [(1 - \Lambda_0(t_i) e^{x_i' \beta})^2 - \Lambda_0(t_i) e^{x_i' \beta}] \\ &= 1 - 3\Lambda_0(t_i) e^{x_i' \beta} + \Lambda_0(t_i)^2 e^{2x_i' \beta} \end{aligned}$$

Specializing to the exponential model, that $\Lambda_0(t_i) = t_i$ leaves us with no additional nuisance parameter in the baseline hazard function. The score function for β is,

$$\nabla_{\beta} \log f(t_i | x_i) |_{\xi=0} = (1 - t_i e^{x_i' \beta}) x_i.$$

The regression coefficient in the residual score for ξ is found to be $\mathbf{a}^T = [-1, 0, \dots, 0]$, which leads to the residual score to be of the form,

$$g(t_i, \beta) = (1 - 3t_i e^{x_i' \beta} + t_i^2 e^{2x_i' \beta}) + (1 - t_i e^{x_i' \beta}).$$

Variance of $g(T_i, \beta)$ can be calculated easily by noting that under the null, $\mathbb{E}(q^k) = \Gamma(k+1)$ where $q = \Lambda_0(T_i) e^{x_i' \beta}$. Since $\mathbb{E}(q^k) = \int q^k f(t_i | x_i) dt_i = \int q^k e^{-q} dq = \Gamma(k+1)$. For the exponential model, $\mathbb{V}(g(T_i, \beta)) = 4$.

The computation for the Weibull model is more involving because of the additional nuisance parameter α . The respective scores for all the parameters are

$$\begin{aligned}\nabla_{\xi}^2 \log f(\mathbf{t}_i | \mathbf{x}_i) &= (1 - \mathbf{t}_i^{\alpha} e^{\mathbf{x}_i' \beta})^2 - \mathbf{t}_i^{\alpha} e^{\mathbf{x}_i' \beta} \\ \nabla_{\beta} \log f(\mathbf{t}_i | \mathbf{x}_i) &= (1 - \mathbf{t}_i^{\alpha} e^{\mathbf{x}_i' \beta}) \mathbf{x}_i \\ \nabla_{\alpha} \log f(\mathbf{t}_i | \mathbf{x}_i) &= \frac{1}{\alpha} + \log \mathbf{t}_i (1 - \mathbf{t}_i^{\alpha} e^{\mathbf{x}_i' \beta})\end{aligned}$$

Let $\theta = (\beta, \alpha)$, the information matrix for the nuisance parameters is

$$I_{\theta\theta} = \sum_i \begin{bmatrix} \frac{\mathbf{x}_i \mathbf{x}_i'}{\alpha} & \frac{\psi(2) - \mathbf{x}_i' \beta}{\alpha^2} \mathbf{x}_i \\ \frac{\psi(2) - \mathbf{x}_i' \beta}{\alpha} \mathbf{x}_i' & \frac{1 - \psi'(2) - 2\psi(2)\mathbf{x}_i' \beta + (\mathbf{x}_i' \beta)^2}{\alpha^2} \end{bmatrix}$$

and the inverse of this matrix is

$$I_{\theta\theta}^{-1} = \sum_i \begin{bmatrix} \frac{\mathbf{q} + (\psi(2) - \mathbf{x}_i' \beta)^2}{\mathbf{q}} (\mathbf{x}_i \mathbf{x}_i')^{-1} & -\frac{\alpha(\psi(2) - \mathbf{x}_i' \beta)}{\mathbf{q}} (\mathbf{x}_i \mathbf{x}_i')^{-1} \mathbf{x}_i \\ -\frac{\alpha(\psi(2) - \mathbf{x}_i' \beta)}{\mathbf{q}} \mathbf{x}_i' (\mathbf{x}_i \mathbf{x}_i')^{-1} & \frac{\alpha^2}{\mathbf{q}} \end{bmatrix}$$

with $\mathbf{q} = 1 + \psi'(2) - (\psi(2))^2$. We further find,

$$I_{\xi\theta} = \sum_i \begin{bmatrix} -\mathbf{x}_i' & \frac{-2 - \psi(2) + \mathbf{x}_i' \beta}{\alpha} \end{bmatrix}$$

The regression coefficient in the residual score for ξ is hence,

$$\mathbf{a}^{\top} = I_{\xi\theta} I_{\theta\theta}^{-1} = \left[\frac{-\mathbf{q} + 2(\psi(2) - \mathbf{x}_i' \beta)}{\mathbf{q}} \sum_i (\mathbf{x}_i \mathbf{x}_i')^{-1} \sum_i \mathbf{x}_i' \quad \frac{-2\alpha}{\mathbf{q}} \right]$$

and $\mathbb{V}(\sum_i g(\mathbf{T}_i, \beta, \alpha)) = \sum_i \mathbb{V}(\nabla_{\xi}^2 \log f) - I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi} = \mathbf{n}(4 - 4/\mathbf{q})$.

B.2. Joint test for Gaussian panel data model. The information matrix for $(\xi, \theta) = (\xi_1, \xi_2, \mu_0, \sigma_0^2)$ is

$$I = \begin{pmatrix} I_{\xi\xi} & I_{\xi\theta} \\ I_{\theta\xi} & I_{\theta\theta} \end{pmatrix} = \frac{NT}{\sigma_0^4} \begin{pmatrix} 2T & \sigma_0^2 & 0 & 1 \\ \sigma_0^2 & (T+3)\sigma_0^4/2 & 0 & \sigma_0^2/2 \\ 0 & 0 & \sigma_0^2 & 0 \\ 1 & \sigma_0^2/2 & 0 & 1/2 \end{pmatrix}$$

We further find

$$I_{\xi,\theta} = I_{\xi\xi} - I_{\xi\theta} I_{\theta\theta}^{-1} I_{\theta\xi} = \begin{pmatrix} 2NT(T-1)/\sigma_0^4 & 0 \\ 0 & NT(T/2+1) \end{pmatrix}$$

and

$$I_{\xi\theta} I_{\theta\theta}^{-1} = \begin{pmatrix} 0 & 2 \\ 0 & \sigma_0^2 \end{pmatrix}.$$

As we have remarked in Section 2.5, the diagonality of $I_{\xi,\theta}$ provides much convenience for finding the optimal test statistics. Denote

$$\mathbf{T}_{\mathbf{n}} := \begin{pmatrix} \mathbf{t}_{1\mathbf{n}} \\ \mathbf{t}_{2\mathbf{n}} \end{pmatrix} = I_{\xi,\theta}^{-1/2} \begin{pmatrix} \sum_i \mathbf{v}_{i1} - 2 \sum_i \mathbf{v}_{4i} \\ \sum_i \mathbf{v}_{2i} - \sigma_0^2 \sum_i \mathbf{v}_{4i} \end{pmatrix} = \begin{pmatrix} (2NT(T-1)/\sigma_0^4)^{-1/2} \left(\sum_i (\frac{\bar{\mathbf{y}}_{i\cdot} - \mu_0}{\sigma_0^2/T} - NT/\sigma_0^2) \right) \\ (NT(T/2+1))^{-1/2} \left(\sum_i (Z_i - T/2)^2 - NT/2 \right) \end{pmatrix}$$

Replacing (μ_0, σ_0^2) by their MLEs yields the joint $C(\alpha)$ test.

APPENDIX C. CLAIM IN SECTION 4

Here we provide the detail derivation for the claim in Section 4 that the reparameterization adopted in Chesher (1984) and Cox (1983) for heterogeneity test requires extra moment conditions on \mathbf{U} for second derivative of log density with respect to the test parameter to be bounded.

Proposition 1. For iid random variable Y_1, \dots, Y_n each with density function $\int p(\mathbf{y}; \lambda_0 + \tau\sqrt{\eta}\mathbf{u}_i)dF(\mathbf{u}_i)$, where \mathbf{U}_i is a random variable with zero mean and unit variance. The second-order derivative of the log density with respect to η evaluated under $\eta = 0$ is unbounded unless $\mathbb{E}(\mathbf{U}^3) = 0$ and $\mathbb{E}(\mathbf{U}^4) < \infty$.

Proof Denote the log density as $l = \log \int p(\mathbf{y}; \lambda_0 + \tau\sqrt{\eta}\mathbf{u}_i)dF(\mathbf{u}_i)$. The first order derivative with respect to η is

$$\nabla_{\eta} l|_{\eta=0} = \frac{\tau \int \nabla_{\lambda} p(\mathbf{y}; \lambda_0) \mathbf{u} dF(\mathbf{u})}{2\sqrt{\eta} \int p(\mathbf{y}; \lambda_0) dF(\mathbf{u})} = \frac{\tau^2}{2} \mathbb{E}(\mathbf{U}^2) \frac{\nabla_{\lambda}^2 p(\mathbf{y}; \lambda_0)}{p(\mathbf{y}; \lambda_0)}$$

The last step is obtained by applying the l'Hôpital's rule.

The second order derivative is

$$\begin{aligned} \nabla_{\eta}^2 l|_{\eta=0} &= \frac{\tau^2 \sqrt{\eta} \int \nabla_{\lambda}^2 p(\mathbf{y}; \lambda_0) \mathbf{u}^2 dF(\mathbf{u}) - \tau \int \nabla_{\lambda} p(\mathbf{y}; \lambda_0) \mathbf{u} dF(\mathbf{u})}{4\eta \sqrt{\eta} \int p(\mathbf{y}; \lambda_0) dF(\mathbf{u})} \Big|_{\eta=0} - \left(\nabla_{\eta} l|_{\eta=0} \right)^2 \\ &= \frac{\tau^3 \int \nabla_{\lambda}^3 p(\mathbf{y}; \lambda_0) \mathbf{u}^3 dF(\mathbf{u})}{12\sqrt{\eta} \int p(\mathbf{y}; \lambda_0) dF(\mathbf{u})} \Big|_{\eta=0} - \left(\nabla_{\eta} l|_{\eta=0} \right)^2 \end{aligned}$$

Provided that $\nabla_{\lambda}^3 p(\mathbf{y}; \lambda_0)$ is not degenerately zero, $\nabla_{\eta}^2 l$ is unbounded unless $\mathbb{E}(\mathbf{U}^3) = 0$ and $\mathbb{E}(\mathbf{U}^4) < \infty$ so that we can apply l'Hôpital's rule again and get

$$\nabla_{\eta}^2 l|_{\eta=0} = \frac{\tau^4}{12} \left[\mathbb{E}(\mathbf{U}^4) \frac{\nabla_{\lambda}^4 p(\mathbf{y}; \lambda_0)}{p(\mathbf{y}; \lambda_0)} - 3\mathbb{E}(\mathbf{U}^2)^2 \frac{\nabla_{\lambda}^2 p(\mathbf{y}; \lambda_0)}{p(\mathbf{y}; \lambda_0)} \right] < \infty$$

■